



WHITE PAPER



FortiSwitch™ Platforms and the High Performance Linpack Benchmark

Results show performance and ease-of-use advantages for 10-Gigabit Ethernet interconnects within HPC environments

Overview

A common misconception of 10 Gigabit Ethernet (GE) in the high performance computing (HPC) community is that it is of lesser performance when compared with other interconnects such as InfiniBand™ and Myrinet™. To be fair, this reputation has not evolved without good reason. Externally generated results for low-level benchmarks such as the Lawrence Livermore National Labs (LLNL) Presta and Intel's Pallas have shown InfiniBand, in particular, to be faster than 10 GE in switchless testing and much faster when operating over a switch fabric. It should be acknowledged that, architecturally, the previous generation of 10 GE switches from companies such as Foundry, Force10, and Cisco, were based on store-and-forward architectures and geared toward enterprise style bulk data transport, focusing on maximizing bandwidth of large packet transfers at the expense of latency in small packets. It is not surprising that the latency penalty for small packet transfers, on this older generation of 10 GE switches, is substantial.

However, the latest generation of Ethernet switches from companies like Fortinet is addressing these issues, and is now able to fully support the demands of the HPC market. Even with these low-level benchmark results, the notion of the fundamental unsuitability of 10 GE for HPC continues to persist. The aim of this technical note is to expose the essential falsity of this position, demonstrate the imprudence of using low-level results as a primary guide to the performance of a cluster system, and to highlight 10 GE as a viable technology for HPC use. The only objective way to do this is through a systematic test of 10 GE as an HPC interconnect technology and comparing to alternatives such as Gigabit Ethernet, which is the most widespread solution. (Gigabit Ethernet was used in 54% of the TOP500 supercomputers in the November 2007 ranking.)

By using the HPC community's most prominent benchmark, the High Performance Linpack (HPL), the relative performance of interconnects can be compared.

This paper establishes that using 10 GE delivers superior performance to Gigabit Ethernet as cluster size scales and achieves equivalent performance to that of InfiniBand or Myrinet. HPL benchmark test results for 10 GE are also on par with alternative interconnect technologies. For the cluster designer accustomed to a default choice of InfiniBand or Myrinet as an interconnect solution, this result highlights a new perspective. Given 10 GE's fundamental compatibility with the computer industry's most widespread networking technology, there are clear advantages to adopting it for the construction of clusters, if no performance penalties are incurred such as those incurred by running native TCP/IP over InfiniBand.

It is now possible, so to speak, to have one's cake and eat it, too. All the benefits of a ubiquitous technology are available to the cluster architect, and with the exception of a few well-understood cases, none of the disadvantages.

Fortinet FortiSwitch- 1000 10-Gigabit Ethernet Fabric Switch

Gigabit Ethernet is considered a poor solution for capability oriented HPC applications for a number of reasons, all related to a lack of performance and scalability. Specifically, the main issues are low bandwidth/high latency for Message Passing Interface (MPI) communications, and more importantly, the tendency for the overall performance of a large Gigabit Ethernet network to degrade badly under load.

By upgrading to a 10 GE solution, most of the bandwidth and latency issues can be addressed, particularly if a low-level protocol stack such as Open Fabrics is used. However, it is also important to consider the issues of scalability and switch saturation due to heavy traffic loads that are a fundamental property of almost all Ethernet installations.

The Fortinet FortiSwitch-1000 10 GE Fabric Switch addresses these issues. Instead of the limitation of single path routing imposed by the use of the spanning tree algorithm, Fortinet uses its vScale™ packet processing technology to build multistage switching fabrics or multipath routed networks. Furthermore, by monitoring one-way latencies within the fabric, the vScale Dynamic Congestion Avoidance feature continuously monitors the current state of the fabric, detects conditions leading to congestion, and actively reroutes and load balances traffic away from congested paths in the fabric to less busy ones.

Fortinet's Dynamic Congestion Avoidance feature operates at a high rate to maintain the lowest latency on all paths. Within a single 144-port FortiSwitch-1000 switch, the highest switch latency between any two ports is 1.6 μ Sec. Fortinet switches can be combined to form large single fabrics to support more than 144 10 GE ports, where the highest latency between any two points on that fabric rises to only 6.4 μ Sec.

High Performance Linpack Benchmark

The High Performance Linpack (HPL) benchmark is probably the single most studied and scrutinized benchmark in the High Performance Computing community. As the benchmark used for the TOP500 list, produced semiannually by the University of Mannheim, University of Tennessee, and NERSC/Lawrence Berkeley National Labs, HPL characterizes the performance of the world's fastest supercomputers and evaluates the performance of newly constructed systems under controlled conditions.

The HPL benchmark measures the performance of a distributed dense linear algebra solver. The result is a single number, R_{\max} , which is the maximum floating-point performance that the solver can deliver from the system being tested.

Note that but for a few exceptional cases, the R_{\max} number actually represents the practical maximum deliverable performance for almost any application run on the system being tested. If the applications user knows how to relate the HPL benchmark to the application being run, then R_{\max} is a better guide than is peak performance in showing how a large single instance of that application will run on the system. This is because the properties of the HPC interconnect are implicitly taken into account with R_{\max} .

When reporting the results of the TOP500 benchmark, the theoretical peak floating-point performance of the system, R_{peak} , is also calculated and included.

An important quantity derived from the HPL benchmark is the ratio of R_{\max} to R_{peak} , a measure that is known as the efficiency of the system. This value may be used in a number of ways. For example, to evaluate the effectiveness of an interconnect, one simply compares the efficiency of two similar system configurations, where the second interconnect changes. The system with the highest efficiency is the most effective.

A more important use of the efficiency number is to determine how well a particular system configuration will scale as problem sizes increase. Because HPL is one of the most efficient benchmarks in the HPC space with regard to making the best use of the architectural features of the system being tested, then any shortcomings in efficiency can be traced directly to those architectural features, such as the cache and memory subsystems, or the interconnect itself.

The HPL benchmark itself is designed to be scalable, both in terms of floating point performance and communications traffic. This means that its efficiency decreases slowly with problem size on a system that is capable of handling the system load required. Therefore, by performing the HPL benchmark at different problem sizes, and calculating the efficiency for each test point, the cluster designer can gauge an accurate notion of the behavior of the system at scale.

Since HPL is a synchronous benchmark, it is sensitive to systemic deficiencies either in the system being tested, or its component subsystems. For example, many clusters have a system-wide monitoring service that is used to collect performance data from all nodes in a cluster. Therefore, when running HPL it is important to turn off those services, because even though these monitoring services impose a small overhead and only slightly slow down HPL on each node, the cumulative effect across the cluster can be dramatic.

Performance Testing

The testing described in this document compares the performance of Message Passing Interface (MPI) communications running over Gigabit Ethernet and 10 GE interconnect technologies. It is critical to be able to perform a true and fair comparison between the interconnects being tested. The key is to limit any changes to a simple reconfiguration or choice of interconnect. By using the HPMPi stack from Hewlett Packard as the MPI layer, we are able to do just that. The HPL benchmark is compiled and built once, and the interconnect to be used for the benchmark run is selected through the invocation of runtime parameters.

We have to be cautious when comparing 10 GE to standard Gigabit Ethernet. Almost all of the major vendors of both 10 GE (and InfiniBand) network devices universally support a low-level protocol stack based on the OpenFabrics RDMA layer, that allows for the interconnect to almost completely bypass the host systems OS layer. This results in a major improvement in performance, both in the raw performance of the interconnect at small and large packet sizes, and because the OS overhead is significantly reduced, allowing a higher throughput. While Gigabit Ethernet device vendors do support zerocopy (i.e., direct to application) protocols, these are usually part of the TCP stack. HPMPi uses RDMA for 10 GE and InfiniBand communications and TCP for Gigabit Ethernet.

The consequence is that while the results obtained are generally close to the best practically available results, it should be possible to do significantly better with Gigabit Ethernet. However, for the purposes of this paper, the results obtained are useful for calibrating the 10 GE results.

It should also be noted that in order to save time, only the 36-node test case was specifically tuned for peak performance, and that the same HPL.dat file derived from that exercise was used in all other test cases, with only the node count and matrix sizes changing. The matrix sizes used for all the cases up to 32 nodes were an estimate based on the amount of available memory for the node count specified and are shown in Table 1. (The reader is directed to Appendix B for more information on what these parameters mean.)

P	Q	P*Q	n
6	6	36	2600
5	6	30	2340
4	8	32	2420
4	7	28	2280
4	6	24	2000
5	5	25	2000
4	5	20	1900
3	6	18	1800
4	4	16	1700
3	4	12	1480
2	5	10	1340
2	4	8	1200
2	3	6	1000
2	2	4	840
1	2	2	600

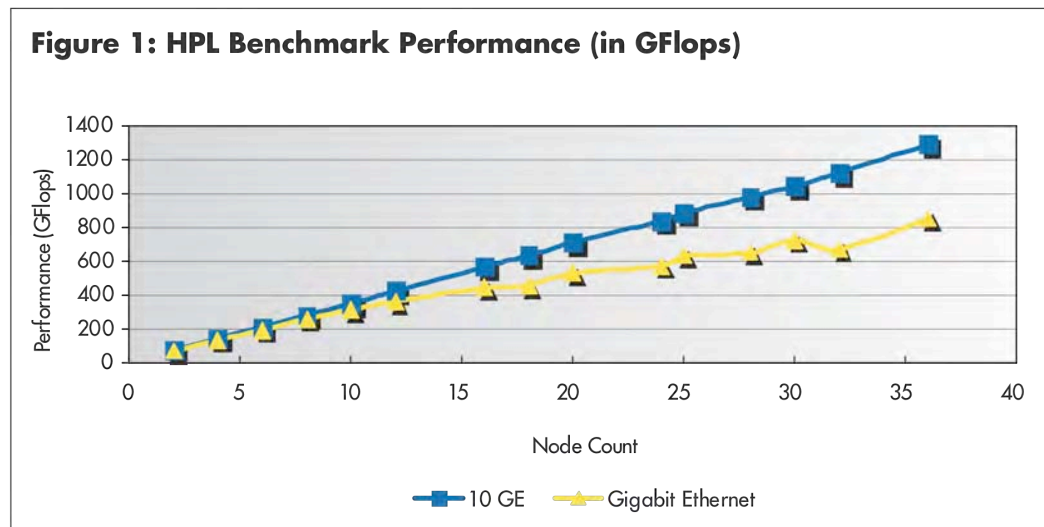
Table 1: Node and Matrix Size Parameters for HPL Benchmark Instances (See Appendix B)

For the dense matrix solver used by HPL, the standard Goto BLAS library is used. This library is generally regarded in the HPC community as the solver of choice for running HPL. The benchmark itself is run in a hybrid mode, and from an MPI point of view, the benchmark is parallel at the node level. Within each node, a multithreaded solver is used, taking advantage of all the cores inside.

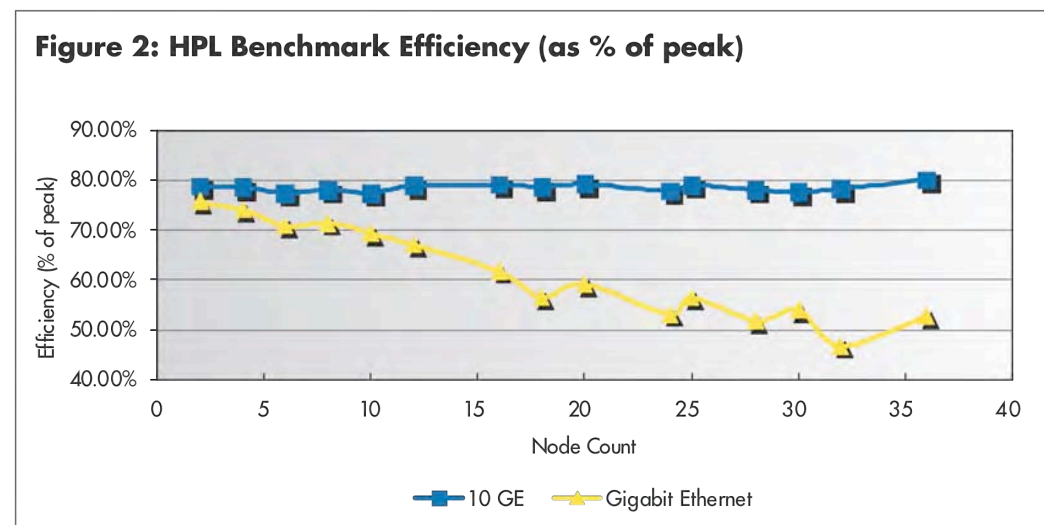
The reason is clear: running in this fashion is the best way to achieve optimal performance with the system configuration at hand. By eliminating MPI-based communications within a node, and constraining any MPI communications to only occur between nodes, the ability to isolate the effects of the interconnect on benchmark performance is greatly enhanced.

Furthermore, each node in this system contains eight processor cores, which is architecturally very similar to nodes in clusters built using server systems based on the new generation of quad-core CPU technology. Therefore, the results outlined here may be easily extrapolated and applied to the latest state-of-the-art systems where a similar model for running HPL is already being followed.

HPL Benchmark Results



Observe how 10 GE follow a straight line that does not fall off. Notice that Gigabit Ethernet, on the other hand, starts to decline at a small node count.



Note that the efficiency for 10 GE is close to a constant 80% through the range of the tests. Contrast that with Gigabit Ethernet's efficiency drop off.

Analyzing the results

Bandwidth performance numbers

From an analytical point of view, the place to start is with the bandwidth performance results for this benchmark. An immediate conclusion of these results is the linearity of the performance numbers for 10 GE, and well within the range of variances in results that we would expect from running these tests many times on the test cluster. In fact, the 10 GE results are on par with results seen in similar tests using double data rate (DDR) InfiniBand. Moreover, we can also discern that the interconnect does matter for this benchmark, as can be clearly seen by the results for standard Gigabit Ethernet, and the way they fall away from the results for 10 GE.

Independently derived, basic raw benchmarks for latency and bandwidth (10 GE's 6.0 μ Secs vs. DDR InfiniBand's 2.2 μ Secs for LLNL's laten latency test, and 1210 Mbytes/sec vs. 1450 Mbytes/sec for LLNL's com bandwidth test) would suggest that DDR InfiniBand should deliver clearly better performance for HPL. Given that the HPL benchmark is affected more by the available bandwidth of an interconnect, than its latency, then the raw numbers would suggest a 20% advantage for DDR InfiniBand over 10 GE. However, when the efficiency numbers for DDR InfiniBand are extracted from the Nov 2007 TOP500 list, and examined, a significant variance for those numbers is seen when the results for different machines are compared. In fact, when the results for DDR InfiniBand capable machines are compared, only one vendor/site is found to be capable of consistently exceeding the 80% efficiency threshold for HPL.

In order to reconcile these two apparently contradictory data points it is important to understand that the raw benchmark numbers can only act as a crude guide to applications performance, because they measure the best possible performance of the MPI stack, run in an environment streamlined to optimize raw benchmark performance. As such, the presta and laten results fail to account for interactions between the interconnect, the application, and the underlying operating system that arise from the execution of the application.

Furthermore, Presta and Pallas only consider the outer boundaries of the performance envelope for an interconnect, i.e., the lowest latency or the highest bandwidth. For example, the laten benchmark from Presta, which measures ping-pong latency between two nodes, does so for the smallest possible packet size and doesn't consider performance at packet sizes that are typically used in real-world applications. In the more conventional regime, the measured performance numbers for 10 GE are much closer to those for DDR InfiniBand than the raw benchmarks would suggest.

Because of its high computational load, the interaction between computation and inter-processor communications, and the range of data packet sizes it uses for its communications, the HPL benchmark is far closer in behavior to real world applications than either presta or com. When run, the HPL benchmark more accurately reflects side effects due to system interactions, and the influence of other system characteristics including optimizations and efficiencies in the interconnect device driver and other software layers.

Benchmark efficiency results

The efficiency numbers for this benchmark also deserve a close look. The results for Gigabit Ethernet clearly show that the traffic generated during the run of the HPL benchmark is indeed significant and brings about a drop in performance due to congestion and other effects caused by the overloading of the Gigabit Ethernet network. It should be noted that the HPL benchmark is only moderately sensitive to latency issues, except where it reduces bandwidth capacity for network traffic characterized by moderate packet size.

For 10 GE, the efficiency stays flat across the range of test cases, suggesting that the interconnect is more than capable of handling the traffic generated by the HPL benchmark. If the cluster being tested were to be expanded and enlarged, then it is very likely that the efficiency value currently seen would also apply to the new, larger system.

Attention should be paid to the observation that for 10 GE, the HPL benchmark is able to achieve 80% efficiency. This number is notable because it has historically been regarded as the threshold for large-scale "Big Iron" systems, such as Cray, NEC and IBM, which have been used to solve the largest capability-oriented problems. While InfiniBand and Myrinet have been regarded as a viable solution for these systems, it is now clear that 10 GE may also be a solution in this problem space.

Implications for real-world applications

One of the criticisms of the HPL benchmark is, that in almost all cases, it is hard to relate the results generated to real-world applications. Such criticism would be valid if the aim were to use HPL to make detailed predictions or judgments about specific applications. However, the goal of this technical note is not to do that, but instead to talk about the properties of interconnects as highlighted by the HPL benchmark, and to apply them in general terms to applications.

What is deducible from running HPL in the scenario described in this paper is that for regular traffic, with payloads of moderate size, 10 GE delivers performance far exceeding that of Gigabit Ethernet and comparable to alternative high performance interconnects. Furthermore, if one knows how an application performs its MPI communications and understands the associated MPI payload profile (i.e., how much traffic represents packets of a given size), one may infer the potential applicability of 10 GE to that application, using the results from this benchmark. For the cluster designer, running the HPL benchmark can provide a simple sanity check for the applications being calibrated, before embarking on analysis that is more detailed.

Clearly not every application will perform as well on 10 GE as on InfiniBand or Myrinet, but it should be pointed out that the majority of applications in the HPC space do indeed have a payload profile of moderate packet sizes that is not unfavorable to 10 GE, including industry standard applications such as LSDyna, Fluent, Abaqus, and others. A few applications, such as Amber and Gamess, operate in a communications regime dominated by small packets, and for these, other interconnects could be a better solution.

It should be noted that the relative constancy in the efficiency results shown here, across the range of problem sizes, could also be an indication as to how well an application may scale on a given cluster. At the very least, it is possible to state that any barrier to applications scalability is unlikely to come from the interconnect fabric, and that given good communications behavior (i.e., packet sizes don't become too small at scale), an application should scale well on 10 GE. The results for Gigabit Ethernet show the repercussions of using an underpowered interconnect.

The importance of using a Fortinet solution for the construction of large-scale systems warrants attention. While it has been shown that 10 GE is superior to Gigabit Ethernet and on an equal footing with other interconnect technologies for the HPL benchmark, the system used for the testing is only of moderate size, and therefore the effects of network congestion for which Ethernet is notorious are not exposed.

Fortinet vScale technology transparently provides a Dynamic Congestion Avoidance capability, allowing the fabric to overcome the inherent limitations of Spanning Tree and other architectural features of standard Ethernet solutions. Using this technology, HPL can run on a large fabric with efficiency numbers similar to those measured by the testing described in this document, up to the largest node counts, just like InfiniBand.

Conclusion

For large fabrics based on regular 10 GE switches without this technology, even though low latency and cut-through routing are supported, the HPL benchmark and other applications will see performance degradation at scale, similar to the Gigabit Ethernet results, due to those inherent architectural features.

The tests described in this technical note show two things. Fundamentally, they demonstrate that the performance of 10 GE is superior to Gigabit Ethernet and on par with other interconnect technologies in the HPL benchmark. Furthermore, by examining the efficiency results, it is shown that the underlying 10 GE fabric, based on Fortinet switches, imposes no barrier to scalability.

These are compelling results. They imply that where cluster designers may previously have overemphasized the needs of the HPC interconnect and sacrificed overall systems performance to better serve the needs of specific key applications; designers can now safely take a broader view of the requirements of their system and users. Other considerations pertaining to the design and architecture of cluster systems may become more prominent.

Like standard Ethernet, 10 GE fully supports a true plug-and-play capability and greatly simplifies the construction and administration of clusters.

Designing a cluster is difficult, in large part because of the issues associated with deciding how to build the underlying communications infrastructure that the cluster needs in order to be functional. At the very least, the system designer has to take into account the needs of three distinct functional parts: the HPC application's interconnect, the Storage Transport layer, and the System Backbone network. What 10 GE specifically provides is an opportunity to consolidate all three of these communications subsystems into a single physical layer, and this is a major simplification. For example, from a cabling point of view, the resulting reduction in cabling volume and complexity enables building of a cluster in a shorter amount of time and can significantly reduce the time it takes to find and repair a fault.

This advantage can be pressed further when using a Fortinet fabric switch network. The ability to load balance and manage traffic, while providing multipath routing, allows such a system to scale well. When the tendency of InfiniBand networks to experience congestion due to static routing is taken into account, it is clear that a 10 GE solution based on Fortinet's fabric switches is in no way an inferior choice.

The consequences are clear. While in some instances, InfiniBand still performs better; in the general case it is hard to discern a clear performance advantage over 10 GE. When the multiuse capability of 10 GE for storage and system backbone functions are considered, along with the inherent ease of design, construction and maintenance of a cluster, it should be clear to any system architect that they should be using 10 GE as the standard scenario. 10 GE is a safe and natural starting point for new cluster designs, and moving away from it is indicated only when there is a specific and compelling need.

For more information and HPL test data for other interconnect technologies, please contact your Fortinet sales representative.

Appendix A: System Configuration

The tests were performed on the Smith cluster at AMD's Development Center lab in Sunnyvale, CA. This system is a 40-node Linux cluster consisting of the following hardware and software:

Compute Node

- Colfax CX1250, quad socket, dual core, 2.8 GHz Opteron with 16 Gbytes RAM
 - ClusterCorp Rocks 4.2.1
 - Linux kernel version: 2.6.942.0.2.ELsmp
- 10 Gigabit Ethernet
 - Intel (NetEffect) 10 GE network interface cards
 - Driver version: Beta 1e driver
 - Firmware version: 202.6
- OpenFabrics 1.2 RDMA layer
- MPI stack
 - HPMPI 2.0

10 Gigabit Ethernet Switch

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

Appendix B:

HPL.dat input deck for the 360node test case

For completeness, the HPL.dat file used to generate the results outlined in this paper is included. It should be observed that between instances of the benchmark, only 3 values change.

- The matrix size shown on line 5, where the value 260000 is substituted for the appropriate value shown in Table 1.
- The Grid size parameters, P and Q. These define how many nodes will participate in the test, and the specific geometry of the grid. These values are also shown in Table 1.

```

Innovative Computing Laboratory, University of Tennessee
HPL.out      output file name (if any)
6           device out (6=stdout,7=stderr,file)
2           # of problems sizes (N). first case is just a smoke test.
2000 260000 # matrix sizes
1           # of NBS 128      NBS
0           PMAP process mapping (0=Row-,1=Column-major)
1           # of process grids (P x Q)
6           P's
6           Q's
16.0        threshold
1           # of panel fact
2 1 2       PFACTs (0=left, 1=Crout, 2=Right)
1           # of recursive stopping criterium
4 4         NBMINs (>= 1)
1           # of panels in recursion
2           NDIVs 1          # of recursive panel fact.
2 1 2       RFACTs (0=left, 1=Crout, 2=Right)
1           # of broadcast
1 1 2       BCASTs (0=1rg,1=1rM,2=2rg,3=2rM,4=Lng,5=LnM)
1           # of lookahead depth
0           DEPTHS (>=0)
2           SWAP (0=bin-exch,1=long,2=mix)
64          swapping threshold
0           L1 in (0=transposed,1=no-transposed) form
0           U in (0=transposed,1=no-transposed) form
1           Equilibration (0=no,1=yes)
8           memory alignment in double (> 0)

```

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.