TEST REPORT
Tolly.

#215111

February 2015

Commissioned by
Mellanox Technologies, Ltd.

Mellanox SwitchX-2 (SX1036) vs. Broadcom StrataXGS Trident II (Arista DCS-7050QX)

Performance Evaluation

Qualifying Data Center Ethernet Networks with RFC2544 at 40Gbps

EXECUTIVE SUMMARY

The demand for data center network performance continues to grow as multi-tenant, public/private clouds and enterprise workloads require that Ethernet switches deliver higher levels of reliability and guaranteed service level agreements. In this environment, unexpected packet loss is unacceptable. In the past Ethernet switches could easily "pass" RFC2544 with no packet loss and did not exhibit large variances in latency. Today, however, that is not the case with some vendors' high-speed switches.

Designing a switch ASIC which operates at 40GbE or higher rates is a different type of challenge and this may be the reason for this new phenomenon where switches fail to pass the very basic RFC2544 at L2 or L3. Today, with the extensive usage of text and short messages, Web2 and large clouds are seeing increasing portions of very small packets which changes the way the network operates.

Mellanox commissioned Tolly to benchmark the 40 Gigabit Ethernet performance of the Mellanox SwitchX-2 ASIC, implemented in the Mellanox SX1036 switch and compare that to the performance of the Broadcom StrataXGS Trident II ASIC, implemented in the Arista Networks DCS-7050QX switch. The Mellanox solution delivered 40GbE wire-speed layer 2 performance with zero frame loss at all frame sizes tested in tests of up to 36 ports. See Table 1.

THE BOTTOM LINE

The Mellanox SwitchX-2 ASIC delivers:

1. Zero-loss, wire-speed throughput at all frame sizes tested from 64- through 9212-byte jumbo frames compared to up to 20% loss and latency up to 97,980ns for Arista Networks
2. Better latency than the Arista Networks DCS-7050QX at all frame sizes tested, up to 96% lower in one test
3. True cut-through switching, while the Arista Networks runs store & forward for 10GbE-10GbE traffic within the same rack for typical top-of-rack topologies

RFC2544 Frame Loss Results: Mellanox SX1036 vs. Arista DCS-7050QX
Layer 2 Multiple 40GbE Ports 100% Line-rate Throughput Test (Part 1)
(as reported by Ixia IxNetwork)

Frame Size (Bytes)		IMIX (30% 1518-, 70% 64-byte)	64	128	256	512	1024	1280	1518	2176	9212
36 Ports /32 Ports (40GbE) Test	Mellanox (36 ports)	0	0	0	0	0	0	0	0	0	0
	Arista (32 ports)	4.3%	19.9%	0	0	0	0	0	0	0	0

Notes: When measuring equipment performance using an IMIX of packets the performance is assumed to resemble what can be seen in "real-world" data center conditions. See https://www.nanog.org/sites/default/files/tuesday_general_nagaranjan_facebook_6.pdf and http://profiles.murdoch.edu.au/myprofile/david-murray/files/2012/06/internet_measurement_2012.pdf for reference. Transmitting rate: 100% line-rate. Ixia traffic mode for the 36 port/32 port test: port 1 to port 2, port 2 to port 3, ..., port n-1 to port n, port n to port 1. The IMIX traffic has 70% 64-byte frames and 30% 1518-byte frames.

Source: Tolly, January 2015

Table 1

Overview: The Case for Zero-Loss, Low-Latency Performance

A switch has but one job - to move every frame across its ASIC as rapidly as possible. Dropping frames and/or excess latency (delay) can only have a negative impact on the applications that are communicating across the switch. For years, switches running port speeds even as high as 10GbE could forward even the smallest 64-byte frames without loss. As this report shows, that isn't necessarily the case with 40GbE switch ASICs and that such frame loss has the potential to impact a range of applications[1].

Cloud & Web2

According to various research reports[2], cloud and Web2 environments consists of network traffic that has a significant percentage of small frames - a mix of approximately 70% 64-byte and 30% 1518-byte frames.

Storage

Storage in general and, more specifically, software defined storage (SDS) scale-out solutions require predictable low latency and high bandwidth. Cut-through switching provides much lower latency than store-and-forward and, thus, the best network performance possible.

Mellanox vs. Broadcom (Arista)

Where the Mellanox SwitchX-2 ASIC delivered wire-speed, no loss 40 GbE throughput at every single frame size and with a 70/30 mix of 64-/1518-byte frames across 36 ports, the Broadcom based Arista switch lost 19.9% percent of 64-byte frames and 4.3% of the mixed traffic(IMIX) when tested using its maximum of 32 ports. See Table 1.

Mellanox Technologies, Ltd.

SX1036

Data Center 40 Gigabit Ethernet Performance

Tolly. Certified

Tested January 2015

Across all test scenarios, the cut-through latency of Mellanox SwitchX-2 is better than that of the Broadcom solution. See Figure 1.

Additional testing benchmarked the performance when a 40GbE port was split into 4 x 10GbE - a common scenario in top-of-rack (ToR) server environments. Testers found that the Broadcom-based solution functioned in store-and-forward for this scenario rather than in cut-through mode, despite the fact it was configured to work in cut-through mode. This resulted in dramatically higher latency for the

RFC2544 Frame Loss Results: Mellanox SX1036 vs. Arista DCS-7050QX

Layer 2 Multiple 40GbE Ports 100% Line-rate Throughput Test (Part 2)

(as reported by Ixia IxNetwork)

Frame Size (Bytes)		IMIX (30% 1518-, 70% 64-byte)	64	84	85
17 Ports (40GbE) Test	Mellanox	0	0	0	0
	Arista	4.3%	20.0%	0.9%	0

Frame Size (Bytes)		IMIX (30% 1518-, 70% 64-byte)	64	71	72
7 Ports (40GbE) Test	Mellanox	0	0	0	0
	Arista	0	8.6%	0.9%	0

Notes: When measuring equipment performance using an IMIX of packets the performance is assumed to resemble what can be seen in "real-world" data center conditions. See https://www.nanog.org/sites/default/files/tuesday_general_nagaranjan_facebook_6.pdf for reference. Transmitting rate: 100% line-rate. Ixia traffic mode for the 7 ports test: port 1 to port 2, port 2 to port 3, ..., port n-1 to port n, port n to port 1. Ixia traffic mode for the 17 ports test: port 1 to port 2, port 2 to port 3, ..., port n-1 to port n. The IMIX traffic has 70% 64-byte frames and 30% 1518-byte frames.

Source: Tolly, January 2015 Table 2

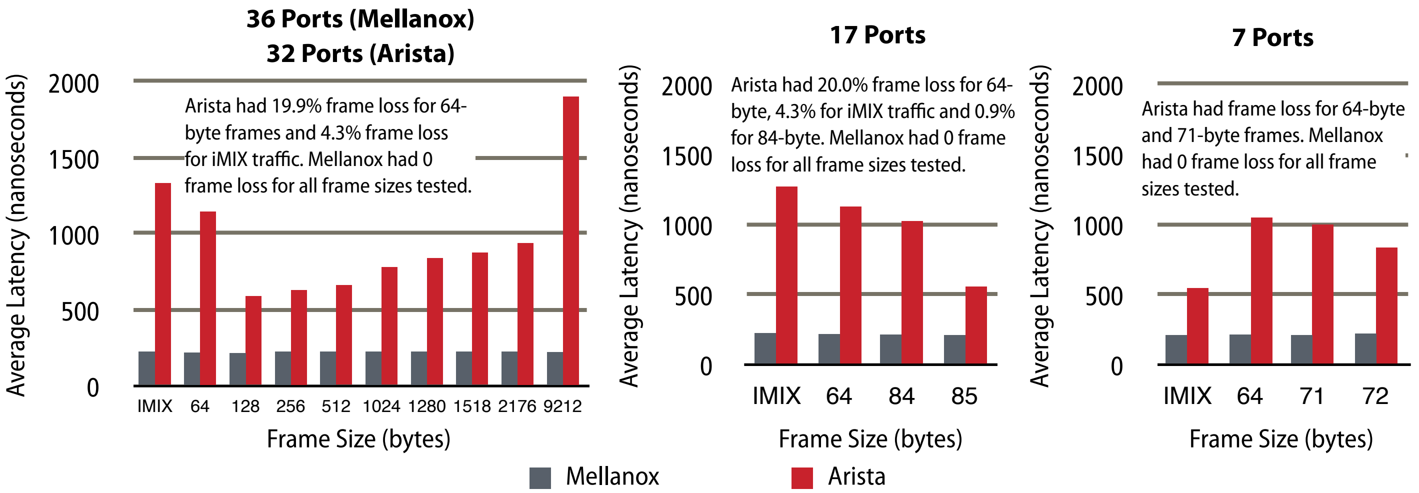
[1] See http://status.ookla.com/incidents/c4c1vyb1ndcr

[2] See https://www.nanog.org/sites/default/files/tuesday_general_nagaranjan_facebook_6.pdf

RFC2544 Cut-through Latency Results: Mellanox SX1036 vs. Arista DCS-7050QX

Layer 2 Multiple 40GbE Ports 100% Throughput Test

(as reported by Ixia IxNetwork)



Arista had 19.9% frame loss for 64-byte frames and 4.3% frame loss for iMIX traffic. Mellanox had 0 frame loss for all frame sizes tested.

Arista had 20.0% frame loss for 64-byte, 4.3% for iMIX traffic and 0.9% for 84-byte. Mellanox had 0 frame loss for all frame sizes tested.

Arista had frame loss for 64-byte and 71-byte frames. Mellanox had 0 frame loss for all frame sizes tested.

Notes: 1. Both switches were configured for cut-through mode. Mellanox SX1036's latency was less than Arista DCS-7050QX's in all tests with the same cut-though forwarding mode. 2. Ixia traffic mode for the 36 ports / 32 ports test and the 7 ports test (chart 1 and 3): port 1 to port 2, port 2 to port 3, ..., port n-1 to port n, port n to port 1. Ixia traffic mode for the 17 ports test (chart 2): port 1 to port 2, port 2 to port 3, ..., port n-1 to port n. The iMIX traffic has 70% 64-byte frames and 30% 1518-byte frames. When measuring equipment performance using an IMIX of packets the performance is assumed to resemble what can be seen in "real-world" conditions. Transmitting rate: 100% line-rate.

Source: Tolly, January 2015

Figure 1

Broadcom solution compared to the Mellanox solution that continued to operate as a cut-through switch. In the worst case of jumbo frames, the Broadcom solution delivered average latency of 7,956 nanoseconds compared to 280 for Mellanox. See Table 3.

Multiple 40GbE Ports Test: Frame Loss and Latency Test Detailed Results

Full System Tests

As noted above, tests were conducted using the full complement of 40 GbE ports for each system which was 36 for the Mellanox solution and 32 for the Broadcom-based Arista solution. The Mellanox solution delivered line-rate throughput at all frame sizes. The Arista solution, as previously noted, showed almost 20% loss with 64-byte frames and 4.3% loss with the mixed traffic. The Mellanox solution had lower latency in all test configurations and up to 88% lower latency in the test of jumbo frames. See Table 3.

17 Ports

To better understand the limitations of the Broadcom solution, Tolly engineers conducted additional testing that reduced both the overall load (i.e., fewer ports) and used different frame sizes between 64- and 128 bytes to attempt to pinpoint the frame size where loss began to occur.

With the overall load reduced to roughly 50% capacity, the performance of both solutions remained consistent. Mellanox again achieved zero-loss while the Arista switch delivered virtually identical loss rates as when 32 ports were tested.

Further testing illustrated that, at this load level, the Arista switch only began forwarding all frames with zero loss when the frame size was 85-bytes. See Table 2.

7 Ports

Engineers further reduced the load to approximately 25% and ran the tests again. As expected Mellanox performance did not change. At this load level, though, the Arista switch was able to forward the IMIX traffic with no loss. Still, frame loss occurred with 64-byte frames with a loss rate of some 8.6%. And, at this load, the Arista switch began forwarding all frames with zero loss when the frame size was 72-bytes. See Table 2.

Two 10GbE Ports Test: Typical 10-40GbE ToR

Network architects can leverage 40GbE ports to connect to multiple servers by splitting a single 40GbE port into 4 links of 10GbE.

Tolly engineers benchmarked a basic server-to-server scenario with 10GbE connectivity to each server. See Figure 3.

As noted earlier, it was in this scenario that engineers observed the Broadcom-based solution actually working in store-and-forward switching mode.

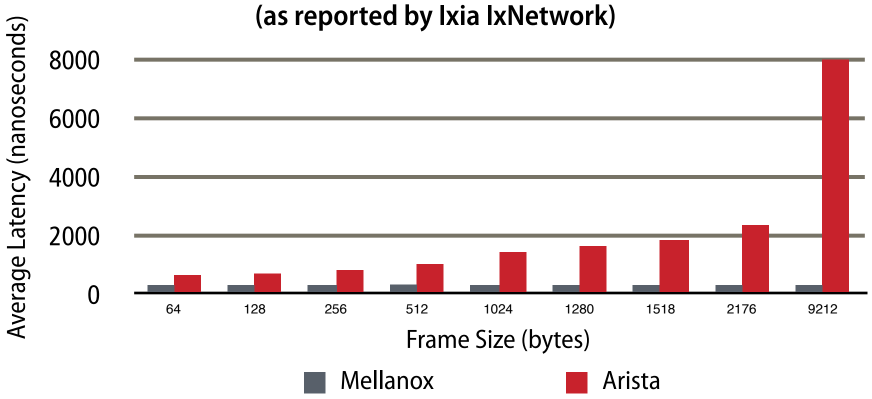
Because of this, the delta in latency between the Mellanox solution, which continued to run as a cut-through switch, and the Arista switch were significantly greater than in prior tests.

For 64-byte frames, Mellanox delivered average latency of 282 nanoseconds compared to 613ns for the Arista switch.

Throughout the range of frame sizes, the Mellanox latency was between 275 and 282ns. With the Arista switch running in store-and-forward mode the latency increased as the frame size increased[3]. Even just with the maximum standard frame size of 1518-bytes this resulted in latency of 1,804ns for the Arista switch compared to 275ns for the Mellanox solution. See Table 3.

Typical 10-40GbE ToR RFC 2544 Latency Results: Mellanox SX1036 vs. Arista DCS-7050QX

Layer 2 Two 10GbE Ports
(as reported by Ixia IxNetwork)



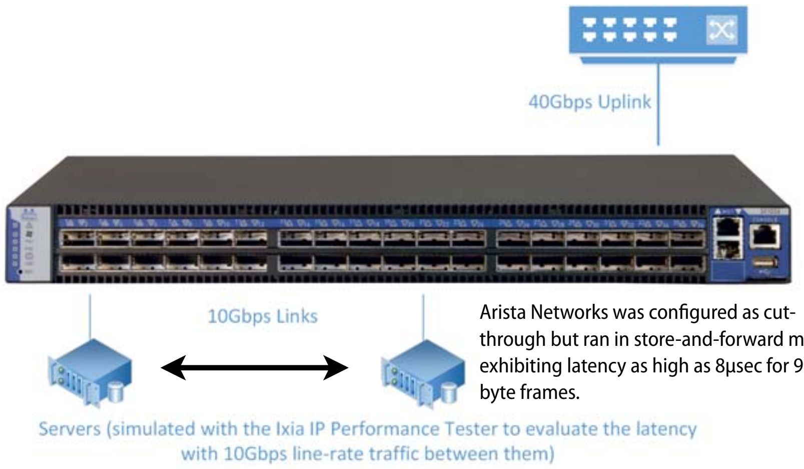
Notes: 1. Both switches were configured to work in cut-through mode. Mellanox SX1036 was actually running cut-through while Arista DCS-7050QX-32-F actually performed store-and-forward switching. Arista DCS-7050QX-32-F supports cut-through mode, however, it appears that the Broadcom Trident II ASIC used in the Arista switch can only run cut-through mode when all ports are running in the same speed. So when administrators are using mixed speeds, which happens in a typical ToR design, the switch can only perform store-and-forward even between ports running the same speed. 2. Neither Mellanox nor Arista experienced frame loss in these tests. 10GbE ports had 100% line-rate traffic. Bidirectional traffic was used in the test. The 10GbE ports under test were split from the 40GbE ports on the switches.

Source: Tolly, January 2015

Figure 2

10GbE Port to 10GbE Port Latency Test Bed

Typical Data Center ToR Switch User Scenario


40Gbps Uplink
10Gbps Links
Servers (simulated with the Ixia IP Performance Tester to evaluate the latency with 10Gbps line-rate traffic between them)

Arista Networks was configured as cut-through but ran in store-and-forward mode exhibiting latency as high as 8μsec for 9212-byte frames.

Note: 10GbE connectivity was achieved through break-out cables.

Source: Tolly, January 2015

Figure 3

[3] The Arista switch was configured as cut-through but the results indicate that it was running in store-and-forward mode.

Mellanox SX1036 vs. Arista DCS-7050QX - All Detailed Layer 2 Latency Results (Nanoseconds)

(as reported by Ixia IxNetwork)

Frame Size (Bytes)			iMIX	64	128	256	512	1024	1280	1518	2176	9212
36 Ports / 32 Ports (40GbE) Test	Average	Mellanox	225	220	216	228	227	227	227	226	226	224
		Arista	1,334*	1,326*	591	631	663	781	840	877	941	1,904
	Maximum	Mellanox	285	320	280	285	292	285	285	282	280	280
		Arista	1,505*	97,980*	692	737	800	945	997	1,040	1,165	2,572
17 Ports (40GbE) Test	Average	Mellanox	224	217	214 for 84-byte frames				222 for 85-byte frames			
		Arista	1,271*	1,129*	1,024 for 84-byte frames*				838 for 85-byte frames			
	Maximum	Mellanox	285	305	252 for 84-byte frames				257 for 85-byte frames			
		Arista	1,372	1,232	1112 for 84-byte frames*				665 for 85-byte frames			
7 Ports (40GbE) Test	Average	Mellanox	213	215	212 for 71-byte frames				222 for 72-byte frames			
		Arista	546	1,051*	1,000 for 71-byte frames*				838 for 72-byte frames			
	Maximum	Mellanox	265	297	252 for 71-byte frames				262 for 72-byte frames			
		Arista	972	1,145*	1097 for 71-byte frames*				652 for 72-byte frames			
One 10GbE Port to One 10GbE Port Test (Typical ToR 10-40GbE)	Average	Mellanox	N/A	282	279	276	285	278	280	275	278	280
		Arista	N/A	613	664	783	996	1,403	1,612	1,804	2,323	7,956
	Maximum	Mellanox	N/A	298	298	298	297	297	295	297	297	294
		Arista	N/A	810	860	980	1,190	1,600	1,810	2,000	2,530	8,150

Notes: *Cells with red text are with frame loss in the RFC2544 100% line-rate test. See the notes of Figure 1 and Figure 2 for other details.
**In the 36 Ports / 32 Ports (40GbE) Test, the 17 Ports (40GbE) Test, and the 7 Ports (40GbE) Test, both switches used the cut-through forwarding mode. In the One 10GbE Port to One 10GbE Port Test, Mellanox SX1036 used cut-through mode. The Arista DCS-7050QX-32-F was configured as cut-through but the results would indicate that the switch was running store-and-forward mode. Arista DCS-7050QX-32-F supports cut-through mode. However, the Broadcom Trident II ASIC used in the Arista switch appears only to run in cut-through mode when all ports are running in the same speed. So when administrators split some 40GbE ports of the Arista switch into 10GbE ports for higher density, the switch appears only to run in store-and-forward mode.

Source: Tolly, January 2015

Table 3

Test Setup & Methodology

Systems Under Test

For Mellanox, the SX1036 switch was tested. This switch had 36 ports of 40GbE and is based on the Mellanox SwitchX-2 ASIC.

For Arista Networks, the DCS-7050QX switch was tested. This switch had 32 ports of 40GbE and is based on the Broadcom Trident II ASIC.

Traffic Generation

All test traffic was generated and all measurements made using Ixia benchmarking equipment consisting of 40GbE test ports in an Ixia XM12 chassis and Ixia IxNetwork 6.30.701.16. Tests were run in port-to-port configuration.

Multiple 40GbE Port Tests

Frame Loss

Engineers chose three configurations to test the switches to see whether the switch under test could support line-rate forwarding. 100% line-rate traffic was sent to the switch under test and the frame loss was recorded.

First, all available 40GbE ports on the switch (36 for the Mellanox SX1036 and 32 for the Arista DCS-7050QX) with traffic topology as port 1 to port 2, port 2 to port 3, ... port n-1 to port n, port n to port 1 were tested. All RFC2544 standard frame sizes were used along with jumbo frames and IMIX traffic which contains 70% 64-byte frames and 30% 1518-byte frames. When measuring equipment performance using an IMIX of packets, the performance is assumed to resemble what can be seen in "real-world" data center conditions. See https://www.nanog.org/sites/default/files/tuesday_general_nagaranjan_facebook_6.pdf and http://profiles.murdoch.edu.au/myprofile/david-murray/files/2012/06/internet_measurement_2012.pdf for reference.

Secondly, 17*40GbE ports on the switch with traffic topology as port 1 to port 2, port 2 to port 3, ... port 16 to port 17.

Thirdly, 7*40GbE ports on the switch with traffic topology as port 1 to port 2, port 2 to port 3, port 6 to port 7, port 7 to port 1.

Port numbers here (e. port 1, port 2, port n, etc.) are used to describe the traffic topology instead of referring to the actual port number. Engineers chose arbitrary ports on the switches to test.

Devices Under Test

Mellanox SX1036 (SX_PPC_M460EX)	Product release: SX_3.4.0250
Arista DCS-7050QX-32-F Switch	Software image version: 4.14.0F

Source: Tolly, January 2015

Table 4

Test Equipment Summary

The Tolly Group gratefully acknowledges the providers of test equipment/software used in this project.

Vendor	Product	Web
Ixia	**Optixia XM12** **Software: IxNetwork 6.30**	ixia TESTED http://www.ixiacom.com

Latency

Latency was measured in the same tests with the throughput tests to compare the cut-through latency of the Mellanox SX1036 and the Arista DCS-7050QX. Tests used Ixia default FIFO latency. All latency results used the Ixia IxNetwork reported latency minus -12ns which is the total latency on the fibers of both sides.

Typical 10-40GbE ToR: Single-Pair of 10GbE Ports

Latency

40GbE ports on both switches can be split into four 10GbE ports for higher density. While traffic passing between the 10GbE ports, the Mellanox SX1036 switch still supported the cut-through forwarding mode. The Arista DCS-7050QX switch, however, automatically worked in store-and-forward mode even though the configuration was still set to cut-through.

The Ixia system was used to simulate two servers to test the latency of this port configuration using bidirectional traffic. No other traffic was passed through the switch under test.

All latency results used the latency reported by Ixia IxNetwork reported minus -12ns which is the total latency on the fibers on both sides.

About Tolly

The Tolly Group companies have been delivering world-class IT services for more than 25 years. Tolly is a leading global provider of third-party validation services for vendors of IT products, components and services.

You can reach the company by E-mail at sales@tolly.com, or by telephone at
+1 561.391.5610.

Visit Tolly on the Internet at:
http://www.tolly.com

Interaction with Competitors

In accordance with Tolly's Fair Testing Charter, Tolly personnel invited representatives from Arista Networks, Inc. to review the test plan and its products results. Tolly did not receive a response to this invitation.

For more information on the
Tolly Fair Testing Charter, visit:

http://www.tolly.com/FTC.aspx

Tolly. Fair Testing Charter

Terms of Usage

This document is provided, free-of-charge, to help you understand whether a given product, technology or service merits additional investigation for your particular needs. Any decision to purchase a product must be based on your own assessment of suitability based on your needs. The document should never be used as a substitute for advice from a qualified IT or business professional. This evaluation was focused on illustrating specific features and/or performance of the product(s) and was conducted under controlled, laboratory conditions. Certain tests may have been tailored to reflect performance under ideal conditions; performance may vary under real-world conditions. Users should run tests based on their own real-world scenarios to validate performance for their own networks.

Reasonable efforts were made to ensure the accuracy of the data contained herein but errors and/or oversights can occur. The test/audit documented herein may also rely on various test tools the accuracy of which is beyond our control. Furthermore, the document relies on certain representations by the sponsor that are beyond our control to verify. Among these is that the software/hardware tested is production or production track and is, or will be, available in equivalent or better form to commercial customers. Accordingly, this document is provided "as is," and Tolly Enterprises, LLC (Tolly) gives no warranty, representation or undertaking, whether express or implied, and accepts no legal responsibility, whether direct or indirect, for the accuracy, completeness, usefulness or suitability of any information contained herein. By reviewing this document, you agree that your use of any information contained herein is at your own risk, and you accept all risks and responsibility for losses, damages, costs and other consequences resulting directly or indirectly from any information or material available on it. Tolly is not responsible for, and you agree to hold Tolly and its related affiliates harmless from any loss, harm, injury or damage resulting from or arising out of your use of or reliance on any of the information provided herein.

Tolly makes no claim as to whether any product or company described herein is suitable for investment. You should obtain your own independent professional advice, whether legal, accounting or otherwise, before proceeding with any investment or project related to any information, products or companies described herein. When foreign translations exist, the English document is considered authoritative. To assure accuracy, only use documents downloaded directly from Tolly.com. No part of any document may be reproduced, in whole or in part, without the specific written permission of Tolly. All trademarks used in the document are owned by their respective owners. You agree not to use any trademark in or as the whole or part of your own trademarks in connection with any activities, products or services which are not ours, or in a manner which may be confusing, misleading or deceptive or in a manner that disparages us or our information, projects or developments.