10GbE Server I/O virtualization

OPENFABRICS ALLIANCE

leonid.grossman@neterion.com

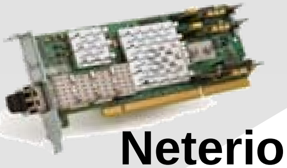
Agenda

- Introduction
- Server I/O virtualization trends and challenges
- 10GbE market trends and challenges
- Server I/0 Virtualization and Sharing with SR and MR IOV 10GbE NICs
- Direct PCI Function assignment to Guest Operating Systems
- Hypervisor assists in Virtual Ethernet Bridges

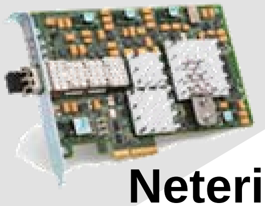
Neterion Introduction

- Shipping 10Gb Ethernet adapters since 2003
- 3rd Gen IOV silicon is shipping now
- Designed for Virtual I/O and Shared I/O Datacenter solutions
- Broad OS and hypervisor support

Neterion Xframe



Neterion Xframe II



Neterion Xframe E

Neterion E3100

Neterion X3100 Series

vmware

Xen

HP-UX 11

FreeBSD

solaris

AIX

Mac

IRIX®

Server I/O virtualization trends and challenges

- Server virtualization is one of the most dominant trends in the datacenter. Virtualization and consolidation are driven by rising datacenter costs of power, real estate and management/support, as well as the multi-core cpu advances.

  In the next few years, most of server growth is expected to be related to virtualization.
- The trend favors smaller number of bigger systems/clusters, as well as more sophisticated I/O adapters in fewer I/O slots.
- I/O virtualization has been lagging cpu/memory virtualization, with most I/O intensive applications remaining on dedicated servers – to a significant degree, due to the lack of IOV solutions with adequate level of performance, isolation and service level guarantees. Such IOV solutions just started to get enabled by advances in hypervisors and arrival of IOV-compliant I/O hardware.
- Virtualization is expected to make a profound impact on Datacenter I/O fabric, accelerating some I/O fabric trends and slowing down other trends.
- At present, the trend may be less applicable to HPC – although the correlation between HPC and broader server market developments is likely to be significant.

10GbE market trends and challenges

- 10GbE adoption in volume servers remains slow relative to earlier Ethernet cycles, mainly due to technology cost/complexity and to the large number of related discontinuities (IOV, converged fabric, rise of multi-core systems).
- 10GbE market remains very fragmented, with large number of vendors playing in different segments and eying eventual “converged fabric” offerings.
- 10G-BT and 10GbE LOM remain a “moving target”, and are not likely to be broadly deployed until cost and power approach GbE levels (no time soon).
- Attach rate for non-virtualized basic 10GbE NICs and blade mezzanine cards remains low, due to the same power/cost constrains.
- Virtual I/O and Shared I/O will likely be the first 10GbE technology competitive with GbE; as such it may become the largest 10GbE segment and to facilitate 10GbE adoption in volume servers.
- Server Virtualization and 10GbE IOV are very complimentary industry developments, and have potential to increase each other’s adoption rate.
- Server I/O virtualization is likely to accelerate some 10GbE NIC features (SR IOV adoption, support for Virtual Ethernet Bridges and other hypervisor hw assists) and somewhat slow down or redefine other features (like iSCSI and RDMA offloads).

Server I/0 virtualization and sharing with SR IOV and MR IOV 10GbE NICs

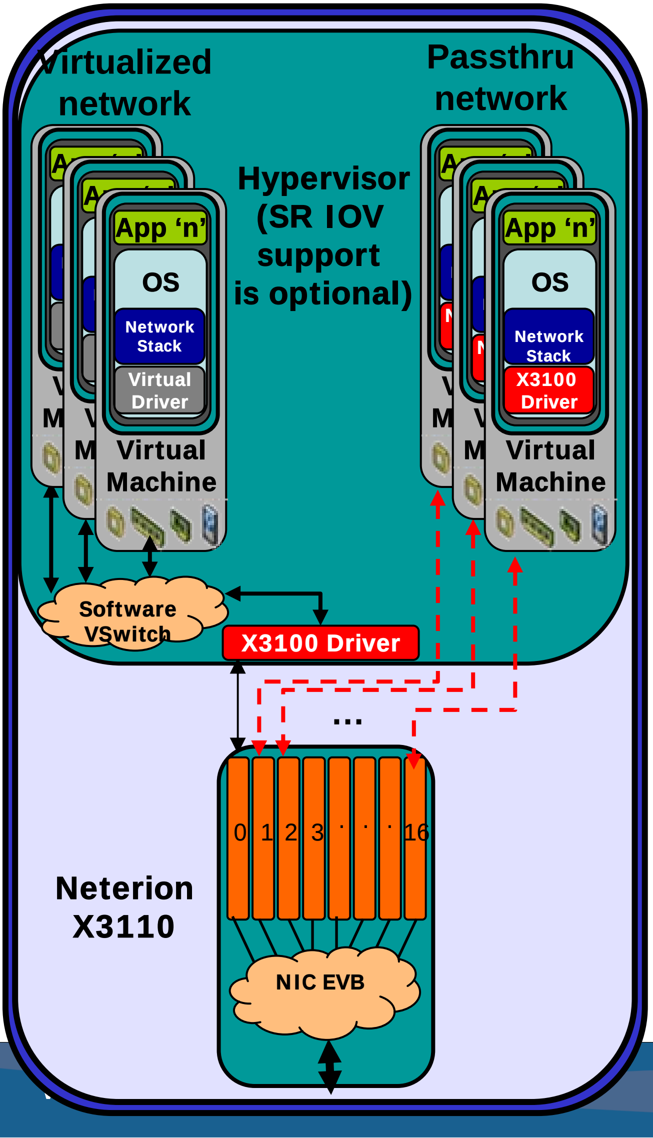
- SR IOV and MR IOV specifications were released by PCI SIG last year.
- SR-IOV compliant NICs are expected to arrive in early 2009 and deliver low-cost support for multiple PCI Functions.
    - From the system view, each function is an independent PCI device (SR IOV OS support will “translate” SR IOV VFs into full PCI functions).
    - SR IOV OS support is “work in progress” in Xen, and is expected to happen in other hypervisors.
    - The functions share 10GbE port(s) and PCI bus; degree of sharing for other NIC resources may vary from vendor to vendor.
    - X3100 hardware is available (including IOV PDK), Linux GPL kernel driver and IOV-related Xen patches are submitted.
- MR IOV capable NICs are expected to facilitate high scalability, low cost/power shared I/O solutions for rack servers and blades. Among other benefits, server administrators will get ability to abstract/pre-configure I/O profiles for both virtualized and non-virtualized servers.

Direct PCI Function assignment to Guest Operating Systems

- Incumbent solutions (emulation and para-virtualization) make progress but still do not meet needs of I/O intensive applications, due to high overhead of I/O virtualization and limited hypervisor ability to provide I/O service level guarantees.
- Running benchmarks at 10GbE line rate in a guest is possible with multi-queue solutions – but at the expense of additional driver complexity and %cpu overhead.
- Direct PCI Function assignment achieves I/O performance similar to native Guest performance on the same system – while preserving virtualization benefits like migration, and leaving control over privileged operations with hypervisors.
- Direct hardware access can be further improved by using “native” driver in a guest OS – same driver is used regardless of which Hypervisor is deployed (or none). This model offers a number of certification and distribution benefits, and supports the same rich native stack feature set in either virtual or native worlds. This model requires SR IOV VF and PF to support the same resources.
- Direct PCI Function assignment can enable virtualization of I/O intensive applications, and increase market penetration for both 10GbE and for hypervisors.

Advanced IOV features supported by IOV 10GbE NICs like X3100

Virtualized Server



In addition to SR IOV and MR IOV support, Neterion X3100 series includes some additional IOV advanced features:

- HW translation of SR IOV to legacy multifunction – direct hw access from a Guest is not gated by SR IOV support in BIOS/OS. X3100 functions work the same way in both IOV-aware and legacy environments.
- PF level support for SR IOV functions. This provides for better protection and isolation that an alternative queue pair based approach. Also,

  no split netfront/netback driver model is required; native Linux or Windows drivers run in guests and see the same set of resources.

  This model allows leveraging work done for the native OS:

  - Driver certification and distribution.
  - Native networking features like Multi-Queue support, Bonding, etc.
- Embedded X3100 10GbE Virtual Ethernet Bridge (VEB) allows a hypervisor to control traditional privileged NIC operations on PCI Functions assigned to guests.

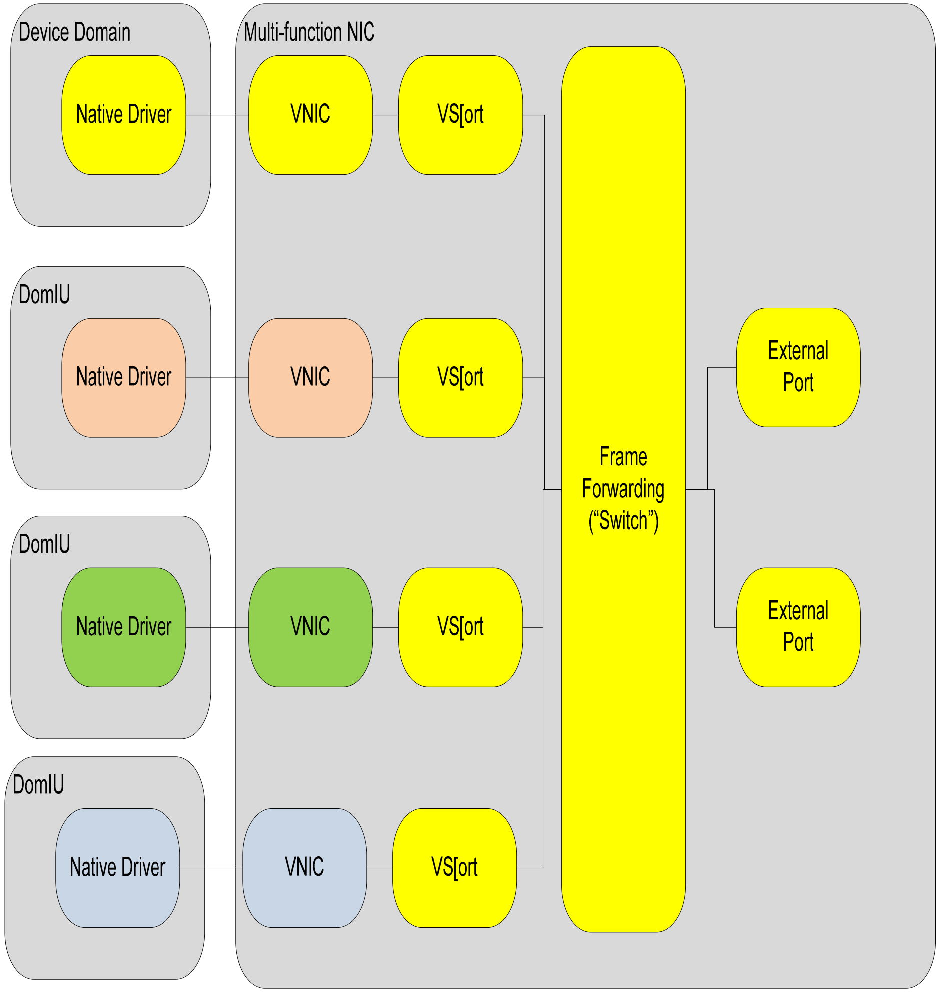
Hypervisor hw assists with Virtual Ethernet Bridges

- VEB is controlled by a privileged domain like a hypervisor or a shared I/O chassis manager. VEB typically supports number of IOV-related capabilities:
  - Direct incoming frames to the correct PCI Function.
  - Support forwarding networking traffic between PCI Functions; Guests can utilize PCI bandwidth instead of 10GbE bandwidth for inter-VM traffic.
  - Allow privileged domain to control traditional privileged NIC operations on PCI Functions assigned to guests. Such control is optional, but entities like Xen Dom0 can control things it chooses, like Guest Vlan membership, etc.
  - Allow an external bridge to control VM connectivity
- Level of NIC VEB compliance with 802.1 bridge specs is vendor-specific.
- An unofficial Edge Virtual Bridging (EVB) Group is recently formed to develop concepts and proposals related to Edge Virtual Bridging for consideration by the IEEE 802.1 working group.
  http://tech.groups.yahoo.com/group/evb/

Support for Direct PCI Function assignment to Guest OS in Xen

- Support for multi-function NICs and function assignment to DomU in Xen exists today, and will get enhanced as Linux support for SR IOV is implemented.
  - Xen supports PCI Function Delegation.
  - Xen supports NIC migration.
  - GOSs support bonding/teaming drivers.
  - GOSs support PCI device insertion/removal.
- An example of configuring multi-function x3100 10GbE NIC for direct access and migration is included later in the slide deck.

A guest controls the Function – NOT the NIC

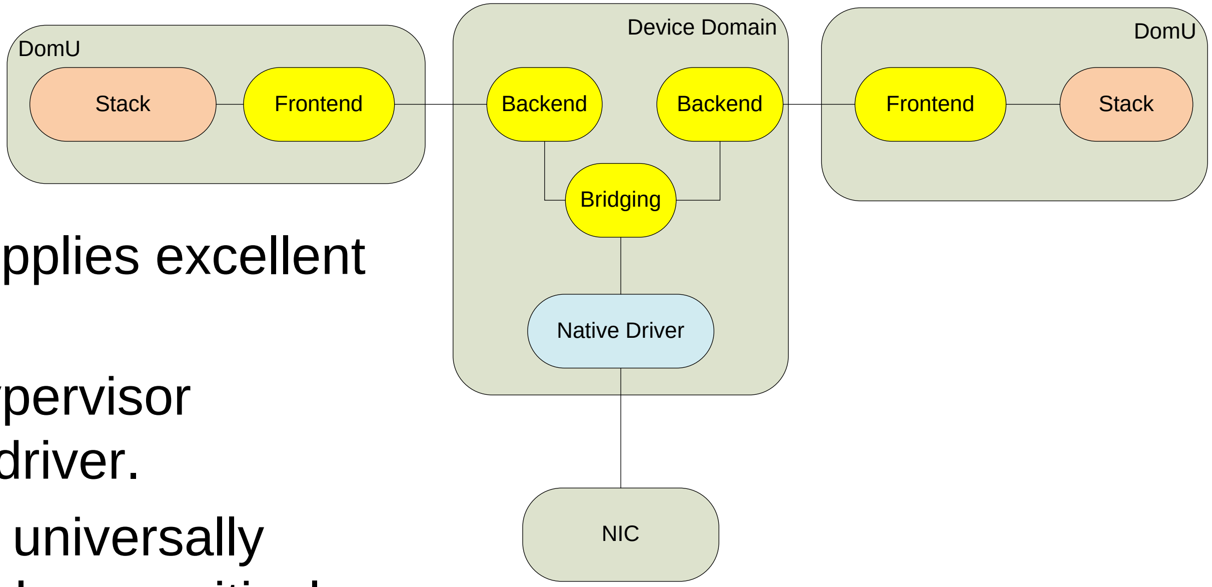


- Each Guest Driver directly controls a VF/VNIC.
- Each Native Driver manages only it’s VNIC.
- Privileged (Device Domain) Driver manages the shared resources.
  - Or whatever portion of them it wants to manage.

Typical resources that Dom0 x3100 driver can choose to manage (via VEB and VF0)

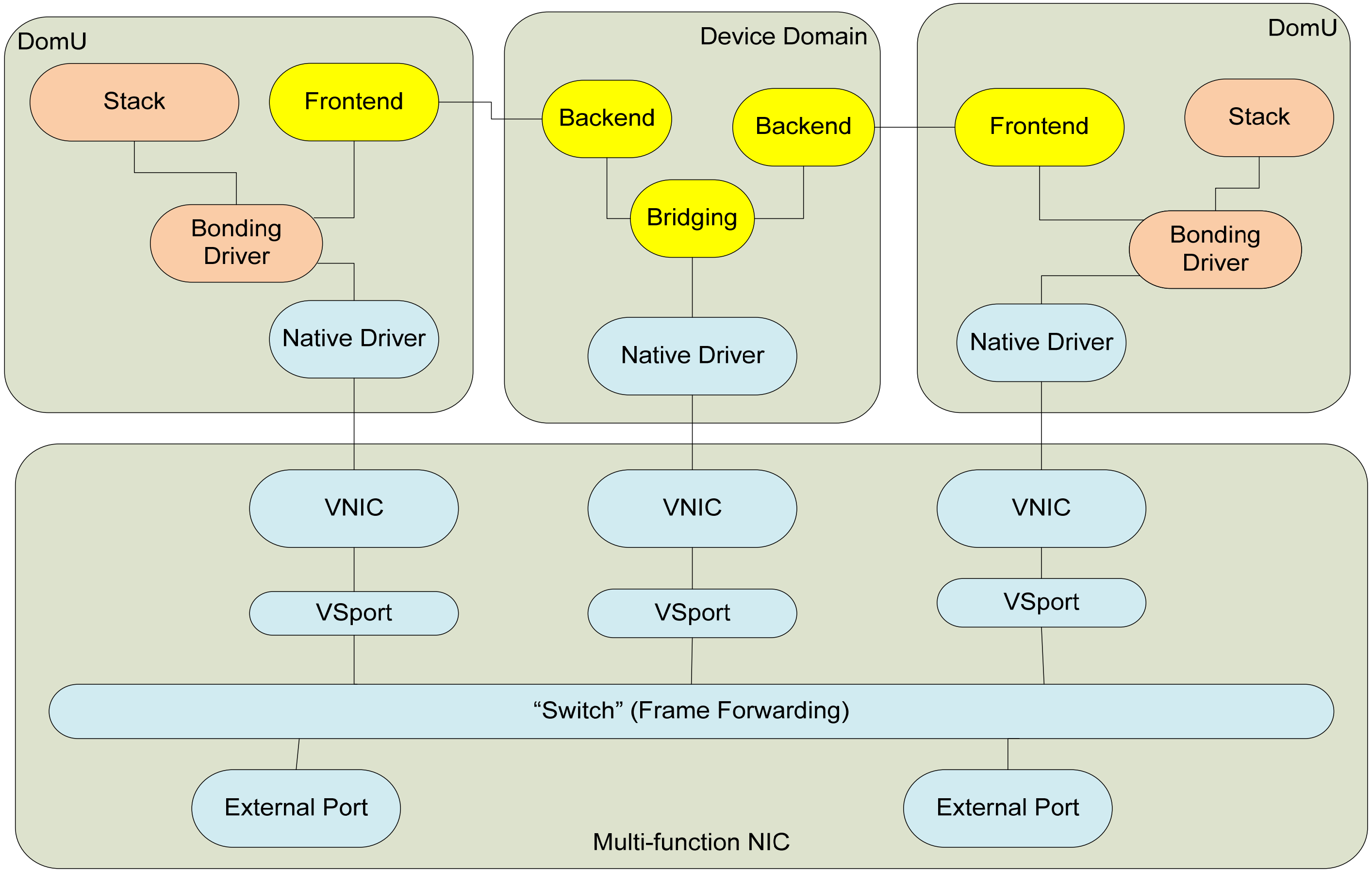
- VF MAC addresses
- VF VLAN membership
- VF promiscuous mode
- Bandwidth allocation between VFs
- Device-wide errors and resets
- VF statistics and FLR
- Link statistics, speed and port aggregation modes

Migration with Frontend/Backend



- Frontend/Backend supplies excellent migration already
  - But requires a Hypervisor specific frontend driver.
- Because it is the only universally supported solution it plays a critical role in enabling migration.

Bonding Direct Assignment with Frontend



Configuring Neterion x3100 for direct access in Xen

- View of X3100 functions in Dom0
  - >>lspci –d 17d5:5833
  - 05:00.0 Ethernet controller: Neterion Inc.: Unknown device 5833 (rev 01)
  - --
  - 05:00.7 Ethernet controller: Neterion Inc.: Unknown device 5833 (rev 01)
- Export a function to DomU, so vxge driver can be loaded:
  - >>echo -n 0000:05:00.1 > /sys/bus/pci/drivers/vxge/unbind
  - >>echo -n 0000:05:00.1 > /sys/bus/pci/drivers/pciback/new_slot
  - >>echo -n 0000:05:00.1 > /sys/bus/pci/drivers/pciback/bind
  - >>xm pci-attach 1 0000:05:00.1
- Configure Bonding interface in DomU with active-backup policy and arp
- link monitoring, so the delegated interface (say eth1) can be enslaved to
- the bonding interface.
  - >>modprobe bonding mode=1 arp_interval=100 arp_ip_target=17.10.10.1
  - >>ifconfig bond0 17.10.10.2 up
  - >>echo +eth1 > /sys/class/net/bond0/bonding/slaves
  - >>echo eth1 > /sys/class/net/bond0/bonding/primary

Live migration for directly attached function

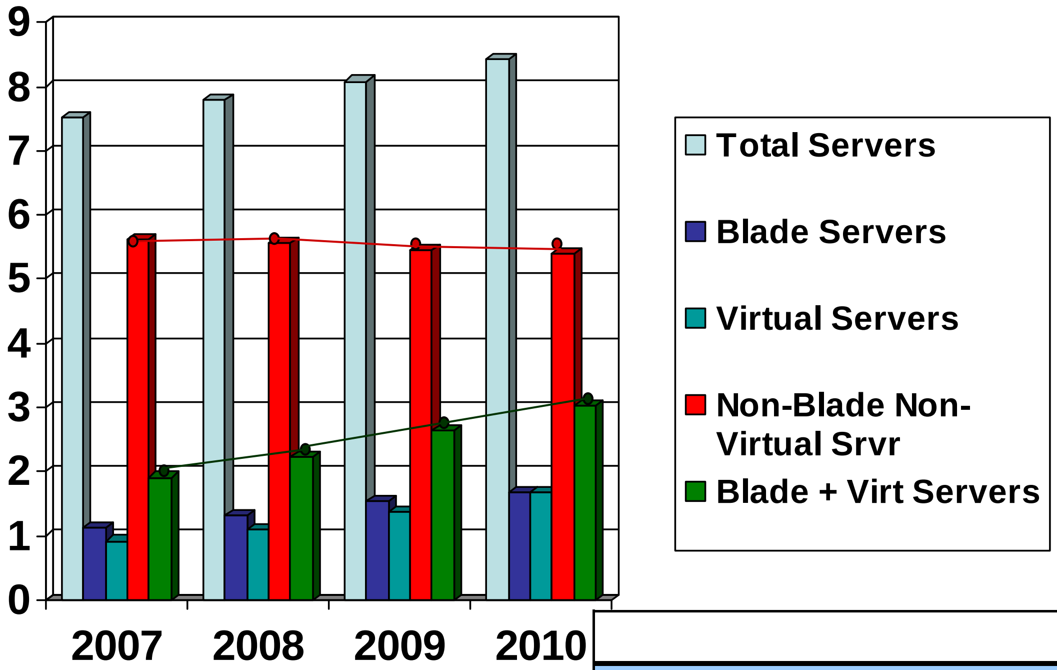
- Create a virtual interface to enslave in DomU as a backup, and remove the delegated interface from DomU to transfer traffic to the virtual interface.
  - `>>xm network-attach 1 bridge=xenbr1`
  - `>>echo +eth2 > /sys/class/net/bond0/bonding/slaves`
  - `>>echo –eth1 > /sys/class/net/bond0/bonding/slaves`
  - `>>ifconfig eth1 down`
  - `>>rmmod vxge`
- Detach pci function from DomU and migrate the DomU to the destination Xen machine (say IP is 172.10.9.7).
  - `>>xm pci-detach 1 0000:05.00.1`
  - `>>xm migrate --live 1 172.10.9.7`
- At this point, network traffic runs on the virtual interface on the destination machine

Moving to direct interface after migration

- Delegate function on the destination machine
  - >>xm pci-attach 1 0000:02:00.1
- Enslave the direct interface
  - echo +eth3 > /sys/class/net/bond0/bonding/slaves
  - echo eth3 > /sys/class/net/bond0/bonding/primary
  - echo -eth2 > /sys/class/net/bond0/bonding/slaves
- Remove virtual backup in DomU
  - >>xm network-list 1
  - >>Idx BE MAC Addr . handle state evt-ch tx-/rx-ring-ref BE-path
  - 0 0 00:16:3e:45:de:53 0 4 8 768 /769 /local/domain/0/backend/vif/1/0
  - 2 0 00:16:3e:61:91:0a 2 4 9 1281 /1280 /local/domain/0/backend/vif/1/2
  - >>xm network-detach 1 2

Backup Slides

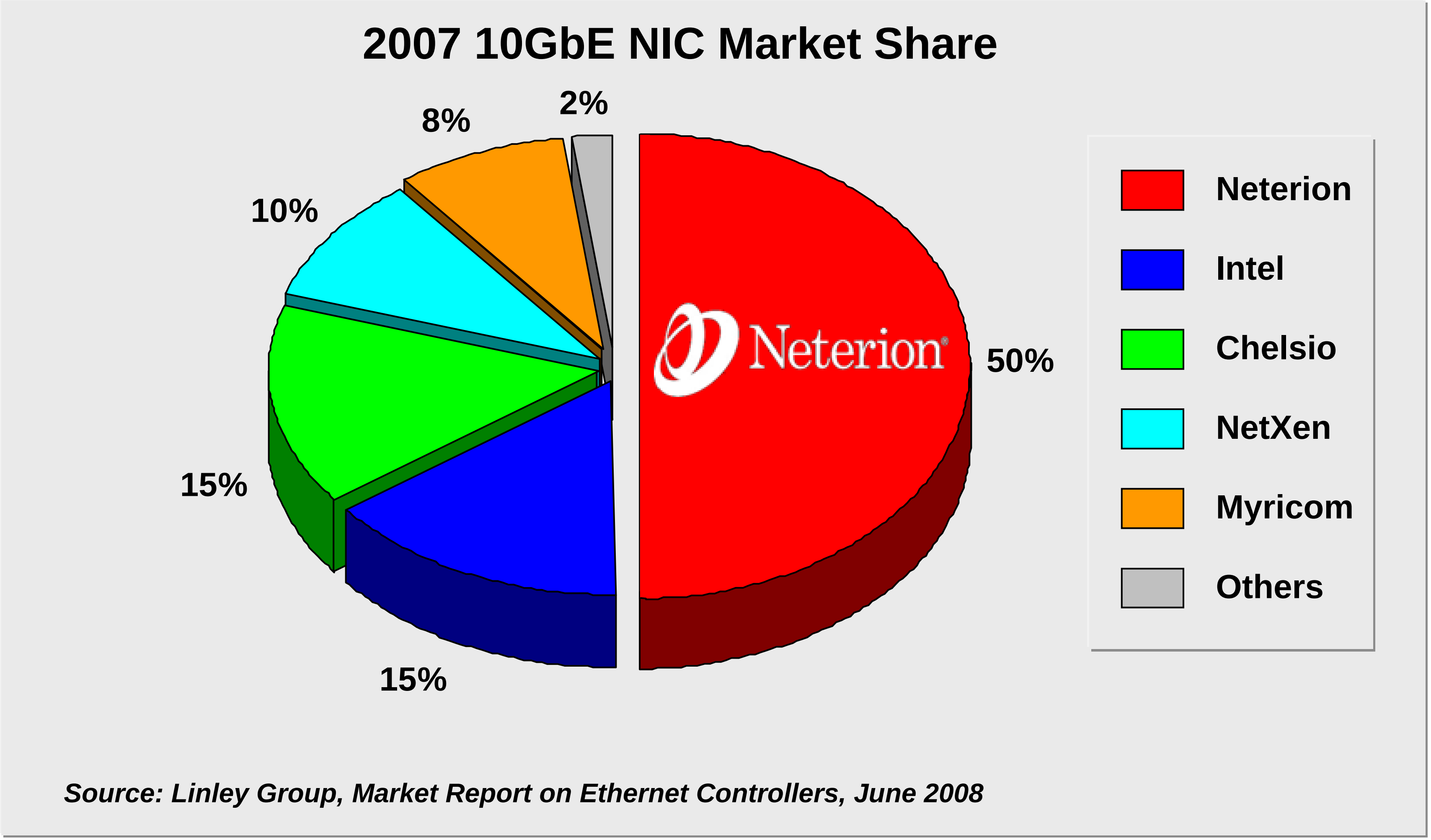
Volume Server (IA) TAM in Units: Servers, Blades, Virtual Servers



- All Server growth is in Blades and Virtual Servers;
- Non-Blade, Non virtual server growth is almost flat

	2007	2008	2009	2010
Total # of Servers (M)	*7.51*	*7.78*	*8.08*	*8.42*
% Virtual Server connect rate	12.2%	14.1%	17.0%	20.0%
% Blade Server connect rate	15.0%	17.0%	19.0%	20.0%
# of Blade Servers (includes Virtual Srvr) (M)	1.13	1.32	1.54	1.68
# of Virtual Servers (includes Blades) (M)	0.92	1.10	1.37	1.68
# of Non-Blade Virtual Servers (M)	0.78	0.91	1.11	1.35
# of Non- Blade Servers (M)	6.38	6.46	6.55	6.74
# of Non-Blade non-Virtual Servers (M)	5.61	5.55	5.44	5.39
# Virtual servers + Blade Servers (M)	1.91	2.23	2.65	3.03

10 GbE NIC Market Share, 2007



Source: Linley Group, Market Report on Ethernet Controllers, June 2008