

Networking Technologies

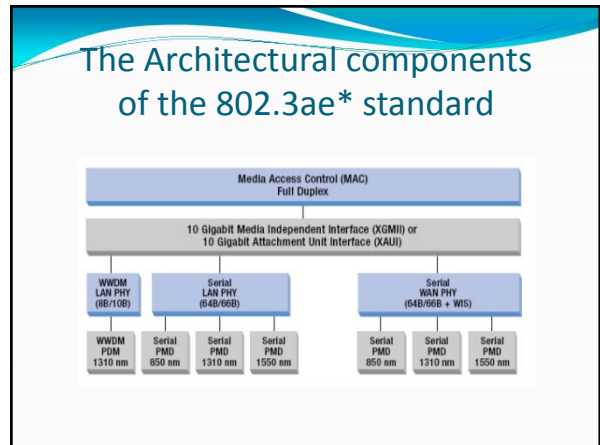
- 10 Gigabit Ethernet
- 40GbE and 100 GbE
- TCP Offload Engine
- iWARP
- RDMAoE

Ethernet's popularity

- 1) Low implementation cost
- 2) Reliability
- 3) Relative simplicity of installation and maintenance

IEEE 802.3ae* standard 10 Gigabit Ethernet Technology

- Significant increase in bandwidth
- Maintaining compatibility with the installed base of 802.3 standard interfaces
- Matches Ethernet model:
 - Media Access Protocol, Ethernet Frame Format, Minimum and Maximum frame size
 - Except it does not need CSMA/CD protocol (Carrier-sensing multiple access with collision detection)



10 Gigabit Ethernet in the Marketplace

- Further distances support – upto 40km
- More bandwidth – 10Gbps

Ethernet now meets the following criteria

- Lower cost of ownership
 - Cabling
 - Equipment
 - Processes and Training
 - Ubiquity of Ethernet
- Familiar management tools and common skills base
- Flexibility in network design – server, switch and router connections
- Multiple vendor sourcing of standards-based products

Applications for 10 Gigabit Ethernet

1. Ethernet as a Fabric Interconnect

- Proprietary networks –
 - Difficult to deploy – experienced IT professionals
 - Higher costs -server adapters and switches
 - Not interoperable with other technologies
- 10 Gigabit Ethernet can replace proprietary technologies
 - Offers the necessary bandwidth
 - Cost saving server consolidation- 7 to 1 savings in management
 - Planned growth of 10 Gigabit network features (eg RDMA/TOE)

2. Local Area Networks

- Can provide better support to the rising number of bandwidth hungry applications
 - Streaming video
 - Medical imaging
 - High-end graphics
 - HDTV
 - Extreme internet gaming
- Can reach greater distances

3. Metropolitan and Storage Applications

- Can build links encircling metropolitan areas with city-wide networks(upto 40km)
- Can support network attached storage(NAS) and storage area networks(SAN)
- Examples of use
 - Business continuance/disaster recovery
 - Remote Backup
 - Storage on demand
 - Streaming Media

3. Metropolitan and Storage Applications

4. Wide Area Networks

Enables ISPs and NSPs to create high speed links at low costs

Need for Higher Speed Ethernet

some critical Internet aggregation points - eight 'lanes' of 10GbE aggregated

Market Trends

- Multi- Core Servers and Virtualization Trend
 - Performance on a Moore's Law curve- doubling every 24 months
 - 40GbE will be the logical next speed for servers in four years
- Networked Storage Trend
 - Disk I/O - primary bandwidth consumers in servers
 - Moving the disks out of the local chassis increases network I/O requirements
 - 40GbE is anticipated to meet the upcoming networked storage bandwidth requirements of the data center

I/O Convergence Trend

- Duplication of hardware infrastructure for local area networks and storage area networks
- I/O convergence demands an increase in the bandwidth requirements
- 40GbE is the next preferred rate

Data Center Network Aggregation Trend

Deployment of 10GbE on servers increases
100GbE proposed to provide the bandwidth to handle the increased traffic load

Graphic courtesy of Google, Inc.

Carrier and Service Provider Networking

- As residential users demand more bandwidth
- The bandwidth requirements of the aggregation of these diverse access networks increases
- 100GbE would be identified as the next high speed network operator interface

Direction for Higher Speed Ethernet

Network Aggregation Bandwidth Increases ~2x/12-18mos. Driven by internet & telecom usage

Computing I/O Bandwidth Increases ~2x/2yrs. Driven by Moore's Law

- 40GbE rate will include a family of physical layer solutions to cover distances inside the data center up to 100m
- 100GbE rate will include distances and media appropriate for data center networking as well as service provider inter-connection for intra-office and inter-office applications

What is TOE and why do we need it?

The Network/System Speed Gap

- The Network/System speed gap is increasing and persisting longer
- A top of the line CPU - fully pegged doing TCP/IP processing at 3 to 4 Gbps of network bandwidth

Dumb NIC Approaches

1. Jumbo Frames

- Ethernet standard limits the frame payload to 1500 bytes
 - at 10Gbps-packet rate 1 million per second
- Jumbo frames decrease the number of packets-processing load on CPUs
- Problems
 - Not standardized
 - Not recognized by previously deployed equipment
 - Not supported by most of the links in the internet
 - Limited on-chip buffering (in today's switch on a chip)
 - Only benefits applications with bulk data transfer

2. TCP Segmentation Offload

- Allows software to pass large TCP packets to the NIC, where they get segmented into standard sizes
- Problems
 - Only helps in large transfers
 - Packet loss results in severe performance degradation

3. Large Receive Offload

- Network Adapter merges incoming packets belonging to the same connection into a larger one
- Problems
 - Delayed Acknowledgments

Arguments against TOE

- TCP Processing is cheap [CLARK89] (1989 study)
- Commodity CPUs scale faster than TOE

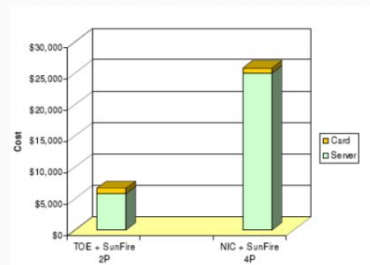
Performance Gains

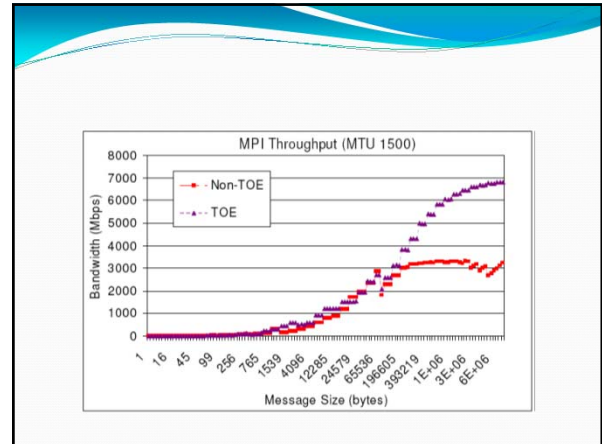
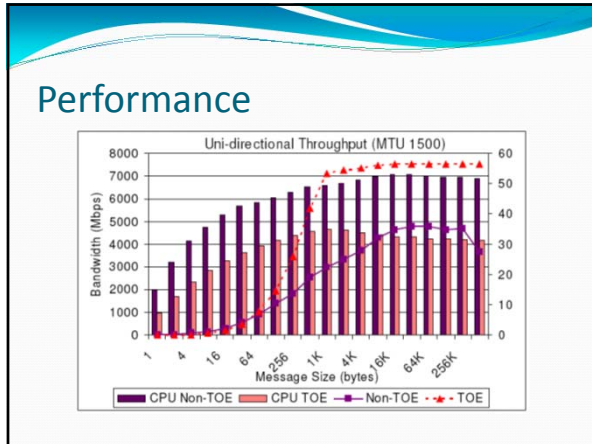
- Direct Data Placement (DDP) – memory subsystem bottleneck problem - receive
- Direct Data Sourcing (DDS) – memory subsystem bottleneck problem - send
- Application layer data integrity check- CRC offload typically used in data critical applications
- Application layer offload- application layer payload recovery, end to end security protocol offload
- Per connection TCP level traffic management and quality of service

Total Cost of Ownership

- Equipment Costs – provide same application level capacity using fewer CPUs and systems
- Management and Development Costs- TOE preserves the popular sockets layer for network programming-eliminates need for dedicated personnel trained in exotic technologies
- Software license fees-eg Database software is typically licensed on per-CPU basis

Cost of ownership





Related Protocols

- RDMA Consortium
 - iWarp
 - iSER (iSCSI Extensions for RDMA)
 - SDP (Sockets Direct Protocol) – applications can gain RDMA benefits without changing their code
- Internet Engineering Task Force (IETF)
 - Many standards related to iWARP
- Microsoft
 - Winsock Direct – RDMA enablement of legacy sockets applications
 - TCP Chimney – Can be used when both servers are not RDMA enabled
- OpenFabrics
 - RDMA acceleration written for MPI
 - RDMA acceleration of popular network storage protocols
 - RDMA acceleration of Linux sockets applications
 - RDMA acceleration of user-level applications via the new OpenFabrics verbs API

RDMA over Ethernet

- Allows running the IB transport protocol using Ethernet frames
- RDMAoE packets are standard Ethernet frames with an IEEE assigned Ethertype, a GRH, unmodified IB transport headers and payload
- InfiniBand HCA takes care of translating InfiniBand addresses to Ethernet addresses and back
- Encodes IP addresses into its GIDs and resolves MAC addresses using the host IP stack
- Use GID's for establishing connections instead of LID's
- No SM/SA, Ethernet management practices are used

