HSSG Speeds and Feeds

Reality Check

Shimon Muller, Andy Bechtolsheim, Ariel Hendel

Sun Microsystems, Inc.

January 2007

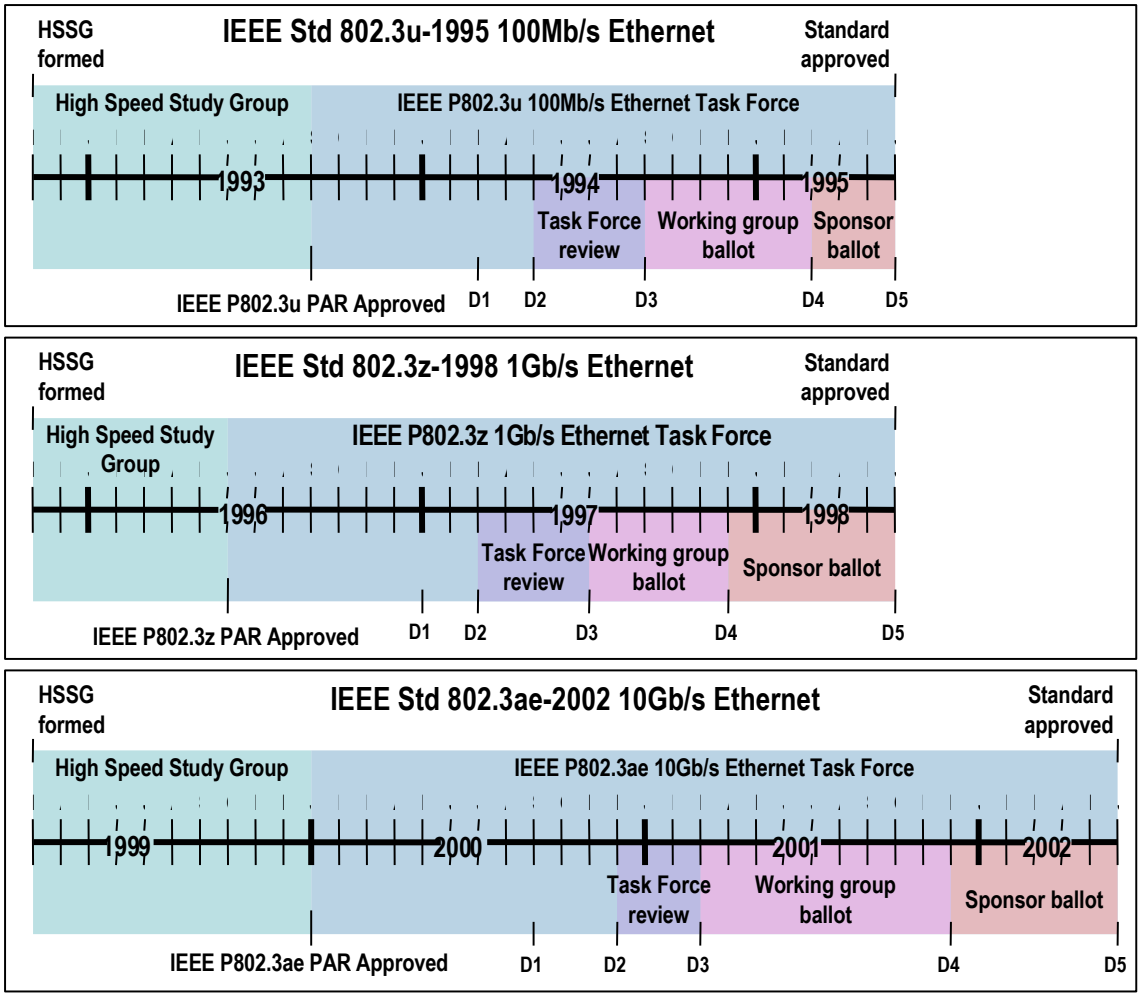
Outline

- Historical Perspective and Trends
  - Lessons learned
- Server Networking Requirements
- HSSG 5-Criteria
- Scaling Ethernet Going Forward
- Things to Avoid
- Summary

SUN: An Early Adopter of Ethernet

- **10 Megabit Ethernet**
  - On every Sun workstation since day one (1982)
- **Fast Ethernet**
  - First to deploy Fast Ethernet on the motherboard (1995)
- **Gigabit Ethernet**
  - First to deploy Gigabit Ethernet on a server motherboard (1998)
- **10 Gigabit Ethernet**
  - Stay tuned...

HSSG History



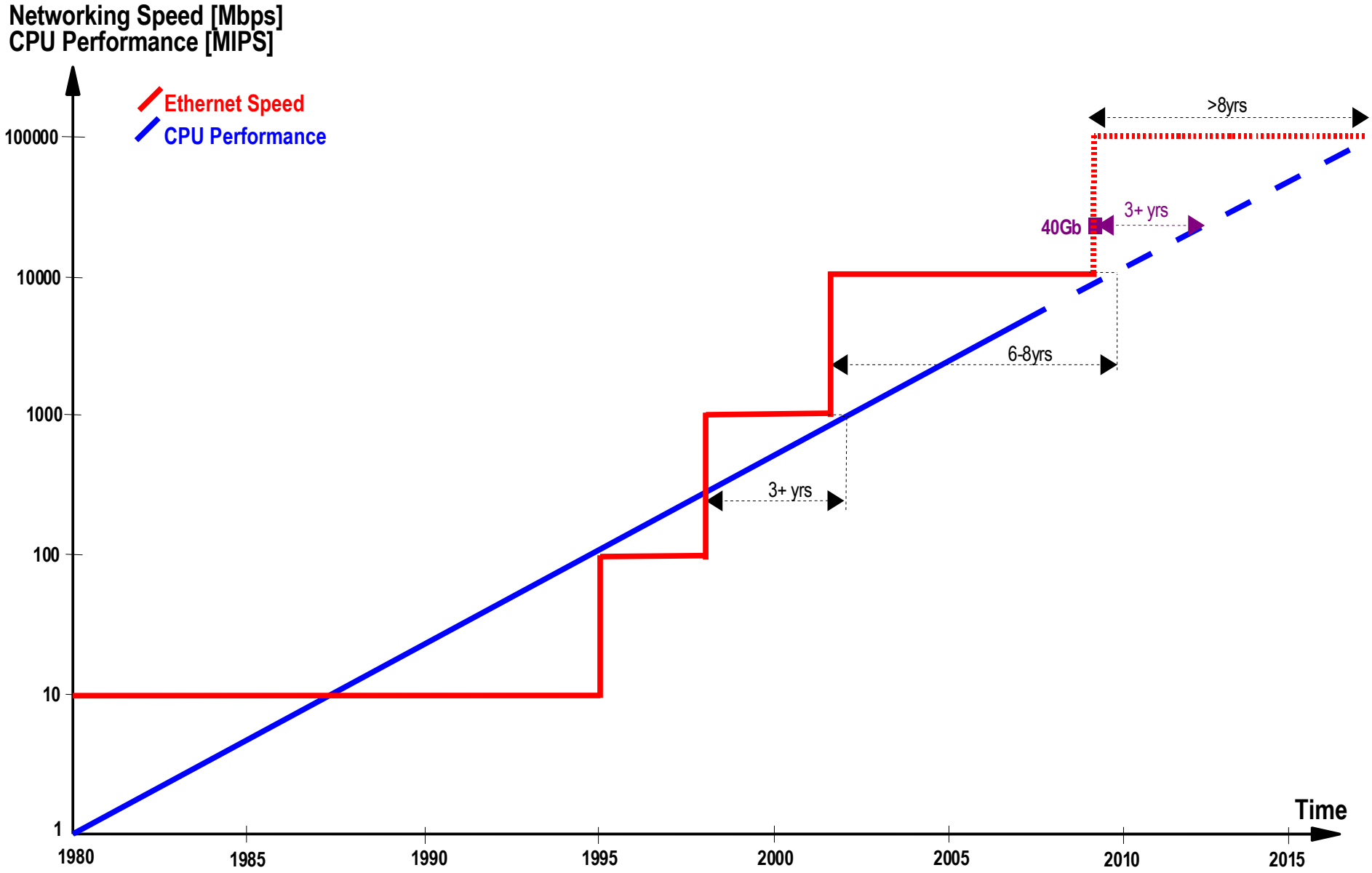
Source: David Law, 3Com

- **Fast Ethernet**
  - 1M host interfaces reached within 1 year from Standard adoption
- **Gigabit Ethernet**
  - 1M host interfaces reached 1-2 years from Standard adoption
- **10 Gigabit Ethernet**
  - 1M host interfaces may be reached 6 years after the Standard was adopted (2008)

Case Study: 10 Gigabit Ethernet

- **Standard Was Ahead Of Its Time**
  - Servers could not support 10Gb/s performance
  - Only market was switch-to-switch backbones
  - The long-haul promise did not pan out
  - Low volume increased cost, leading to slow adoption
- **PHY/PMD Technology Was Not Ready**
  - The cost of the PMD still dominates deployment cost
  - Four generations of optical modules (Xenpack, X2, XFP, SFP+)
  - Still waiting for widespread deployment of 10GBase-T
- **Market Was Not Ready**
  - No pre-standard deployments
  - Computer systems could not utilize the 10Gb/s bandwidth
  - Just having the standard did not create market momentum
  - Volume Adoption will start in 2008, 6 years after standard adoption
- **Lesson Learned: Creating a new networking standard ahead of its time does not accelerate market adoption**

CPU Performance Drives Network Speeds



System Throughput Drives Network Needs

- **System throughput doubles roughly every 2 years**
  - **Implies the following network throughput roadmap:**
    - 10Gbps in 2007
    - 40Gbps in 2011
    - ***100Gbps in 2014 --- a long time from now***
    - 160Gbps in 2015
    - 640Gbps in 2019
- **100 Gigabit Ethernet will not be usable by servers until 2015**
  - Server I/O roadmaps will not support it any time sooner
    - PCI-Ex Gen-2 (DDR) can support no more than 40Gb/s
    - Requires PCI-Ex Gen-3 (QDR)
- **A network technology that cannot connect to computers is (by definition) low volume**
  - Basically limited to inter-switch links and the long-haul
- **Next generation Ethernet *MUST* address a larger market to justify the investment**

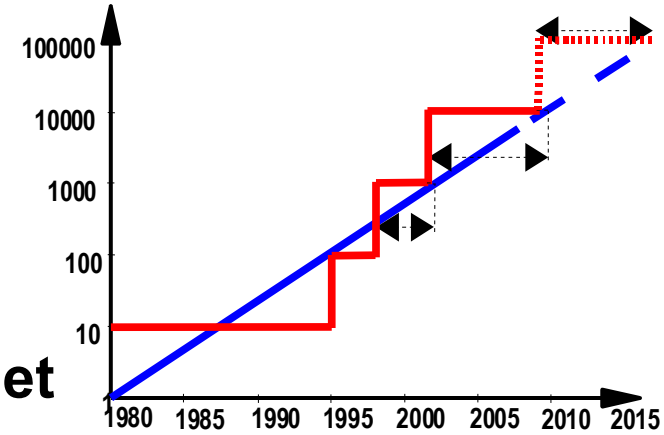
HSSG and the 5-Criteria



- **The Essence of Ethernet is Broad Market Potential**
  - **This tradition is more important than the 10x scaling factor**
    - Every new Ethernet speed has always embraced the broadest market possible
  - **Networks that cannot connect to computers are not very interesting**
    - Servers are where most of the volume will come from
    - Backbones and long-haul by themselves do not drive volume
- **Technical Feasibility is irrelevant unless taken in the context of Broad Market Potential and Economic Feasibility**
  - There are any number of things that can be built
  - What makes sense to build is the question
- **Economic Feasibility *requires* broad market potential, and Broad Market Potential *requires* connectivity to servers**

The 10x Scaling Factor

- **10x scaling worked great from 10Mb to 100Mbit, and then to 1000Mbit**
  - > Technology required was readily available, sometimes leveraged
  - > Hit the volume market more or less on time
- **It did not work well for 10 Gigabit Ethernet**
  - > Had to develop the required technologies
    - > To some extent continues to this day
  - > Long lead time to volume (6+ years)
- **For sure it will not work for 100 Gigabit Ethernet**
  - > Transistor speeds do not jump by 10x every few years
  - > Neither does computer throughput
- **Selecting 100Gbps PHYs and PMDs is a major challenge**
  - > Picking a technical solution today for a 2015 market is a very high risk
  - > Investment needed to productize the technology requires volume markets
    - > Cost of new technology development is escalating, both for silicon and optics



“Scalable Ethernet” Architectural Framework

- **As part of the HSSG effort, define an architectural foundation based on the following principles:**
  - **Flexible enough to scale Ethernet well into the future**
    - More than 10 years
    - Up to 1Tb/s and beyond
  - **Based on existing 802.3 layering model, with necessary enhancements**
    - Speed-independent MAC
    - Multi-lane, multi-rate *n*-GXS and *n*-AUI
    - Multi-lane, multi-rate RS and *n*-MII
  - **Leverage is good --- no need to reinvent the wheel**
    - **PCI-Express already has most of the above defined**
      - 1-lane, 4-lane, 8-lane and 16-lane options
      - The lanes can run at multiple speeds: 2Gb/s, 4Gb/s, ...
      - Auto-negotiation defined for selecting the lane configuration
      - May need to make some changes for Ethernet-specific items

“Scalable Ethernet” Architectural Framework

- **Standardizing a new Ethernet speed becomes easy**
  - Add parameters in speed-specific table(s)
    - i.e. 4.2 in the MAC
  - Define a bit-time
  - Define a lane configuration
  - Define the rate per lane
    - More than one lane/rate configurations may be allowed
  - If needed, define a new PHY and PMD

Next Speed(s)

- **One intermediate speed is needed before we get to 100Gb/s**
  - Do 40Gbps first, as 4x10Gbps
    - Can be deployed in the near future
  - Will help to meet Broad Market Potential for 100Gbps
    - Higher speeds at the edge will drive higher speeds in the core sooner
  - Do 100Gbps as 4x25Gbps, not 8x12.5Gbps or 10x10Gbps
    - Going wider is not the same thing as going faster

- **Long Lead time for 100Gbps PHY/PMD Development**
  - Parallel copper or optics are limited to very short reach
  - Longer reach needs WDM
  - Fewer lasers is always better than more lasers

Things to Avoid

- **Using prime numbers for the number of lanes in parallel interfaces (i.e. 3, 5, 7,...)**
  - There is nothing in the industry today that is x3, x5 or x7
    - Connectors, optical modules, cables, etc.
  - Multiples of prime numbers are highly undesirable
- **At the logical level everything is doable, but may take an unnatural act to accommodate**
  - Like ATM cells of 53 bytes...
- **Computers like powers of 2, so stick with it!!!**
  - Please...

Summary

- **Broad Market Potential and Economic Feasibility require us to consider how the next generation network connects to servers**
  - Server throughput doubles roughly every two years
    - This already accounts for multi-core and multi-threaded processors
    - Clock rate increases will be very modest going forward
- **Increasing network speeds by 10x does not meet server needs**
  - 40Gbps is a much more desirable next step for 2010
    - It will take until 2015 for servers to need 100Gbps
- **100Gbps-only Ethernet will not achieve full market potential**
  - Will remain an expensive backbone solution for a long time
    - Backbones only do not drive volume economics
  - Will delay the adoption of next generation Ethernet
    - Creating a new networking standard ahead of its time will not accelerate market adoption
- **We need a scalable interface below the MAC**
  - Backwards compatible XAUI extension