400GbE and High Performance Computing

John D’Ambrosia

The last time the IEEE 802.3 Working Group addressed the “Next Rate” of Ethernet was when 10 GbE was Ethernet’s fastest rate. That effort resulted in the development of two new rates — 40 GbE and 100 GbE. The justification for two rates was that 40 GbE was intended to provide the upgrade path for servers, while 100 GbE would target network aggregation applications. Since the ratification of the 40 GbE and 100 GbE standard, however, a different scenario has emerged. While the market continues to transition to 10 GbE-based servers from GbE-based servers, 40 GbE is seeing adoption and deployment in data center networks, and 100 GbE is seeing adoption and deployment in a number of data aggregation intensive applications, such as network cores, service provider client connections, internet exchanges and others.

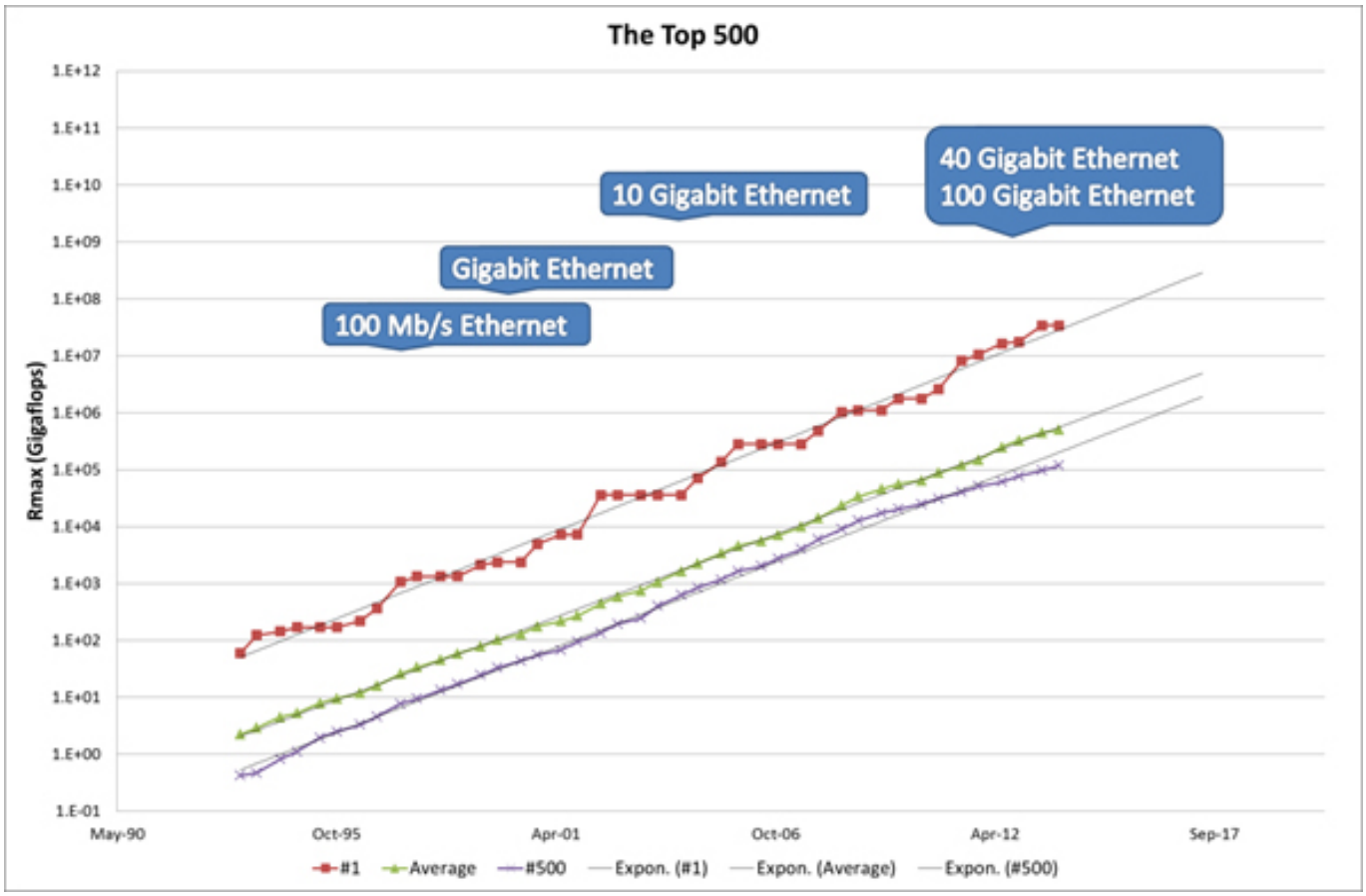
However, the IEEE 802.3 Ethernet Bandwidth Assessment Ad hoc, performed a year-long industry study of bandwidth requirements. Its assessment demonstrated that industry bandwidth requirements are continuing at an exponential pace — growing by a factor of 10 every five years. This translates to, on average, the need to support terabit-per-second capacities in networks by 2015 and 10-terabit-per-second capacities by 2020.[1] Recognizing this industry need, the IEEE 802.3 formed the IEEE 802.3 400 Gb/s Ethernet (400 GbE) Study Group in May 2013. The group defined the project, which will focus on solutions for network aggregation applications, such as cloud-scale data centers, Internet exchanges, co-location services, wireless infrastructure, service provider and operator networks, and video distribution infrastructure. This project was recently approved, and will be addressed by the IEEE P802.3bs 400GbE Task Force, which will meet for the first time at the IEEE 802.3 May 2014 Interim Session to begin work on the definition of the 400 GbE standard.

The objectives of this project target the noted network aggregation application

needs and include:

- Support a MAC data rate of 400 Gb/s
- Support a BER of better than or equal to 10^{-13} at the MAC/PLS service interface (or the frame loss ratio equivalent)
- Support full-duplex operation only
- Preserve the Ethernet frame format utilizing the Ethernet MAC
- Preserve minimum and maximum FrameSize of current Ethernet standard
- Provide appropriate support for OTN
- Specify optional Energy Efficient Ethernet (EEE) capability for 400 Gb/s PHYs
- Support optional 400 Gb/s Attachment Unit Interfaces for chip-to-chip and chip-to-module applications
- Provide physical layer specifications which support link distances of at least
    - 100 m over MMF
    - 500 m over SMF
    - 2 km over SMF
    - 10 km over SMF

The question for the readers of this article, however, is how will 400 GbE be applicable to high performance computing? There are two very simple ways to address this question.



First, let's consider HPC systems as assessed by TOP500.org, which has been releasing their twice-a-year report since June of 1993. Figure 1 plots the Rmax statistic, which is the maximal LINPACK performance achieved, of the #1 and #500 systems, as well as the average of all of the TOP500 systems. Note the trend lines

and the predictable performance.

Systems based on Gigabit Ethernet propelled Ethernet as an interconnect family for the TOP500 going back to 2001, and enjoyed a near-continuous growth for eight years. In recent years, the presence of Gigabit-Ethernet-based systems in the TOP500 has diminished. However, in the past two years, the presence of 10-Gigabit-Ethernet-based-systems has seen a sharp uptick of nearly 650 percent growth.

However, to keep up with the predictable growth of HPC-based systems, it is clear that ever-increasing rates at lower costs will continually need to be introduced. And with Ethernet, there is clearly a roadmap of rates, as well as a never-ending focus on interoperability, that drives completion which, in turn, drives cost reduction.

The need for 400 GbE goes beyond its actual use in the development of systems for the TOP500. As noted by the TOP500 itself, as justification for the list, “These people wish to know not only the number of systems installed, but also the location of the various supercomputers within the high performance computing community and the applications for which a computer system is being used. Such statistics can facilitate the establishment of collaborations, the exchange of data and software, and provide a better understanding of the high performance computer market.”[2]

Collaboration. Bob Metcalfe famously described in Metcalfe’s Law how the value of a network grows exponentially as a function of the number of users. For example, consider the Energy Sciences Network (ESnet), a high-speed computer network serving United States Department of Energy (DOE) scientists and their collaborators worldwide. In a presentation to the IEEE 802.3 Ethernet Bandwidth Assessment Ad hoc in December 2011, there was an expectation that ESnet would need to support 100 Petabytes per month of data in 2015.[3] Connecting a vast network of scientists at over 40 institutions, as well as more than 100 other research and education networks,[4] it should come as no real surprise that a review of the data regarding ESnet accepted traffic shows a 10x growth in data every five years,[5] mirroring exactly the IEEE 802.3 Ethernet Bandwidth Assessment.

So, while it will most likely be some time before finding its way into a large number of HPC systems in the Top500, 400 GbE will play a vital role in the core networking interconnecting these HPC systems!

References

1. IEEE 802.3 Industry Connections Ethernet Bandwidth Assessment, http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf [1], July 2012.

2. Top500 – Introduction and Objectives, http://www.top500.org/project/introduction/ [2].

3. “Data Intensive Science Impact on Networks”, Darney, Eli, http://www.ieee802.org/3/ad_hoc/bwa/public/dec11/dart_01a_1211.pdf [3], Dec 2011.

4. ESnet – Connected Sites, https://www.es.net/overview-of-the-network/connected-sites/ [4].

5. IEEE 802.3 Industry Connections Ethernet Bandwidth Assessment, http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf [1], July 2012

John D'Ambrosia is chairman of the Ethernet Alliance and chief Ethernet evangelist, CTO office at Dell. He may be reached at editor@ScientificComputnig.com [5].

Source URL (retrieved on *09/03/2014 - 1:10am*):
http://www.scientificcomputing.com/blogs/2014/04/400gbe-and-high-performance-computing

Links:
[1] http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf
[2] http://www.top500.org/project/introduction/
[3] http://www.ieee802.org/3/ad_hoc/bwa/public/dec11/dart_01a_1211.pdf
[4] https://www.es.net/overview-of-the-network/connected-sites/
[5] mailto:editor@ScientificComputnig.com