Server Bandwidth Implications for the Next Higher Speed of Ethernet

David Chalupsky, Intel Corporation

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

Geneva, CH

Goal

- Review Server Ethernet connectivity history and forecast to project need for higher speed Ethernet

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

How Will Server Ports Impact HSE?

- Are we in a hurry for a 400Gbit NIC? 1Tbit NIC?
  - No, not this decade…
- But bi-sectional bandwidth is dependent on the links in the aggregation layer.
- "Top of Rack" topologies connect a small number of servers in one rack to the LAN via an *even smaller* number of uplink ports.
- Ratio of total access bandwidth to total uplink bandwidth is the "oversubscription ratio."
  - High oversubscription is bad for applications with high server-to-server communication

> Increase in Server access port speed drives the
> need for higher uplink bandwidth

September 2012

Fat and Flat

- Contemporary applications require more communication between servers than in the past.
  - Search, indexing, databases, technical computing, analytics on "Big Data"
- Ideal network to serve these apps has non-blocking, full BW between all servers
  - Sounds expensive…
  - Let's look for a reasonable compromise…

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

A *Simple* Look at Oversubscription

- Let's take a typical ToR switch…
- 48 access ports, 2-4 uplink ports

Number of Access Ports	Access Port Speed (Gbps)	Number of Uplink Ports	Uplink Port Speed (Gbps)	Total Access BW (Gbps)	Total Uplink BW (Gbps)	Oversubscription Rate
48	1	4	10	48	40	1.2
48	10	4	10	480	40	12.0
48	10	2	40	480	80	6.0
48	10	4	40	480	160	3.0
48	40	4	40	1920	160	12.0
48	40	2	100	1920	200	9.6
48	40	4	100	1920	400	4.8
48	40	2	400	1920	800	2.4
48	40	4	400	1920	1600	1.2
48	100	4	100	4800	400	12.0
48	100	2	400	4800	800	6.0
48	100	4	400	4800	1600	3.0
48	100	2	1000	4800	2000	2.4
48	100	4	1000	4800	4000	1.2

Baseline:
48x1G -> 4x10G =
1.20 OS ratio

…that's really good.

A *Simple* Look at Oversubscription (2)

- Let's take a typical ToR switch...
- 48 access ports, 2-4 uplink ports

Number of Access Ports	Access Port Speed (Gbps)	Number of Uplink Ports	Uplink Port Speed (Gbps)	Total Access BW (Gbps)	Total Uplink BW (Gbps)	Oversubs cription Rate
48	1	4	10	48	40	1.2
48	10	4	10	480	40	12.0
48	10	2	40	480	80	6.0
48	10	4	40	480	160	3.0
48	40	4	40	1920	160	12.0
48	40	2	100	1920	200	9.6
48	40	4	100	1920	400	4.8
48	40	2	400	1920	800	2.4
48	40	4	400	1920	1600	1.2
48	100	4	100	4800	400	12.0
48	100	2	400	4800	800	6.0
48	100	4	400	4800	1600	3.0
48	100	2	1000	4800	2000	2.4
48	100	4	1000	4800	4000	1.2

10G access w/ 10G uplink was horribly oversubscribed.

40G uplinks offer good options for 3.0->6.0 OS for 10G access

A *Simple* Look at Oversubscription (3)

- Let's take a typical ToR switch…
- 48 access ports, 2-4 uplink ports

Number of Access Ports	Access Port Speed (Gbps)	Number of Uplink Ports	Uplink Port Speed (Gbps)	Total Access BW (Gbps)	Total Uplink BW (Gbps)	Oversubscription Rate
48	1	4	10	48	40	1.2
48	10	4	10	480	40	12.0
48	10	2	40	480	80	6.0
48	10	4	40	480	160	3.0
48	40	4	40	1920	160	12.0
48	40	2	100	1920	200	9.6
48	40	4	100	1920	400	4.8
48	40	2	400	1920	800	2.4
48	40	4	400	1920	1600	1.2
48	100	4	100	4800	400	12.0
48	100	2	400	4800	800	6.0
48	100	4	400	4800	1600	3.0
48	100	2	1000	4800	2000	2.4
48	100	4	1000	4800	4000	1.2

40G access needs 100G uplinks… at least.

400G uplinks offer GREAT options for 1.2->2.4 OS for 40G access

A *Simple* Look at Oversubscription (4)

- Let's take a typical ToR switch…
- 48 access ports, 2-4 uplink ports

Number of Access Ports	Access Port Speed (Gbps)	Number of Uplink Ports	Uplink Port Speed (Gbps)	Total Access BW (Gbps)	Total Uplink BW (Gbps)	Oversubscription Rate
48	1	4	10	48	40	1.2
48	10	4	10	480	40	12.0
48	10	2	40	480	80	6.0
48	10	4	40	480	160	3.0
48	40	4	40	1920	160	12.0
48	40	2	100	1920	200	9.6
48	40	4	100	1920	400	4.8
48	40	2	400	1920	800	2.4
48	40	4	400	1920	1600	1.2
48	100	4	100	4800	400	12.0
48	100	2	400	4800	800	6.0
48	100	4	400	4800	1600	3.0
48	100	2	1000	4800	2000	2.4
48	100	4	1000	4800	4000	1.2

100G access with100G uplinks is horribly oversubscribed

400G uplinks offer good options for 3.0->6.0 OS for 100G access

September 2012

Oversubscription Summary

- 40GbE NICs
  - 100Gb/s uplinks: okay for a start: OS=4.8
  - 400Gb/s uplinks: Great! OS=1.2

- 100G NICs:
  - 100Gb/s uplink – miserable
  - 400Gb/s uplink – pretty good, OS=3.0
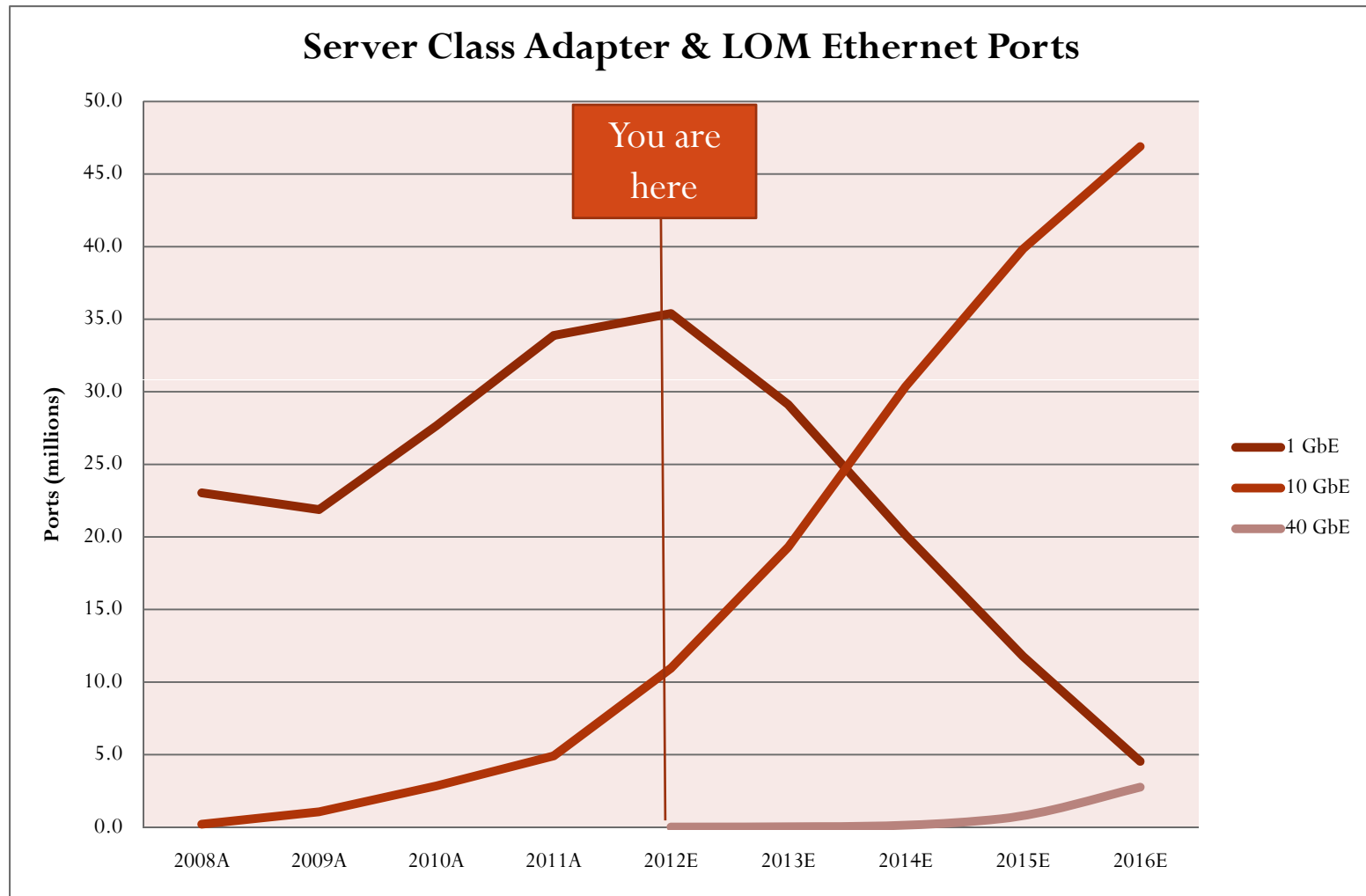 - Terabit uplinks: great… but can we wait?

400Gb/s uplinks support Server access of 40Gb/s and 100G/s …so when will we need it?

The State of Things
in the Server Ethernet Market

- Overall port count growth ~20%/yr, 2008-2012
  - Expected to slow to ~5% 2013+ as higher speed ports deploy
  - Users saw multiple 1G ports as more cost effective than 10G

- Gigabit Ethernet
  - The incumbent technology, with ~76% of the ports in 2012
  - GbE still growing in 2012 …may finally be peaking
    - "Rumors of my death have been greatly exaggerated."

- 10 Gigabit Ethernet
  - On a very strong growth ramp
  - Expected to surpass 1GbE ports in 2014

- 40 Gigabit Ethernet
  - Just getting started, expect to hit 5% of ports in 2016

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

What That Looks Like In Server Ports

Server Class Adapter & LOM Ethernet Ports

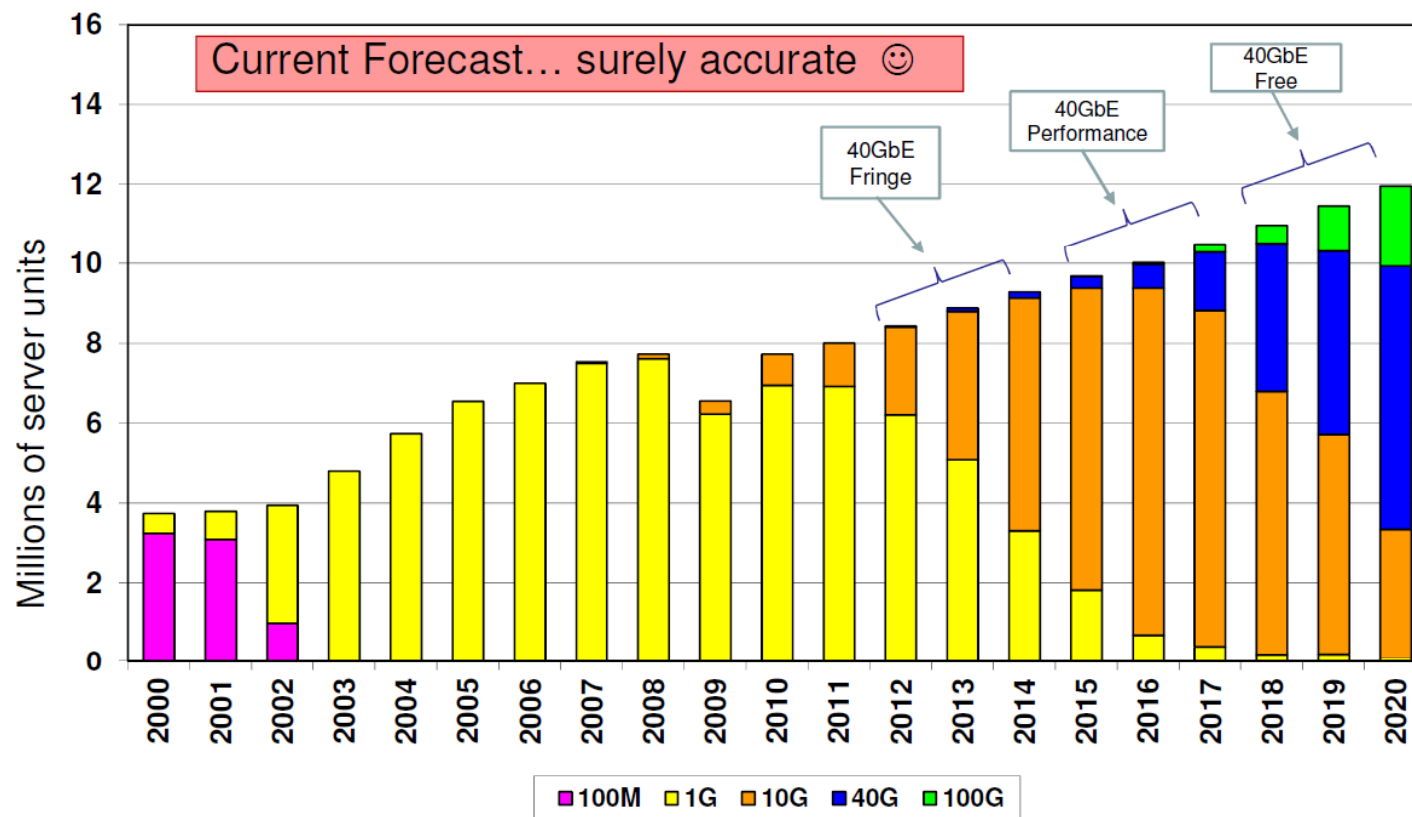Source: Crehan Research, 2012

IEEE 802.3 Higher Speed Ethernet Consensus Ad Hoc

September 2012

From the Next Generation BASE-T CFI

x86 Servers by Ethernet Connection Speed
(2012 Forecast)
Based on IDC, Dell Oro, Crehan Research and Intel data from 2H'11 – 1Q'12

Current Forecast... surely accurate ☺

40GbE Free

40GbE Performance

40GbE Fringe

Millions of server units

Legend: 100M · 1G · 10G · 40G · 100G

Version 1.1   IEEE 802.3 Next Generation BASE-T CFI Consensus Building Presentation – July 2012 Plenary   Page 19

Opinions About the Future Differ

- Key questions for a long term forecast
  - Rate of Server BW growth
  - Rate of higher speed port adoption
  - 40G or 100G NICs?
    - "40G is a small step, I'm waiting for 100G!"
    - "40G will be low cost by using mature technology. I'm going to stick with that for a long time!"

Reliable information about the future is hard to come by
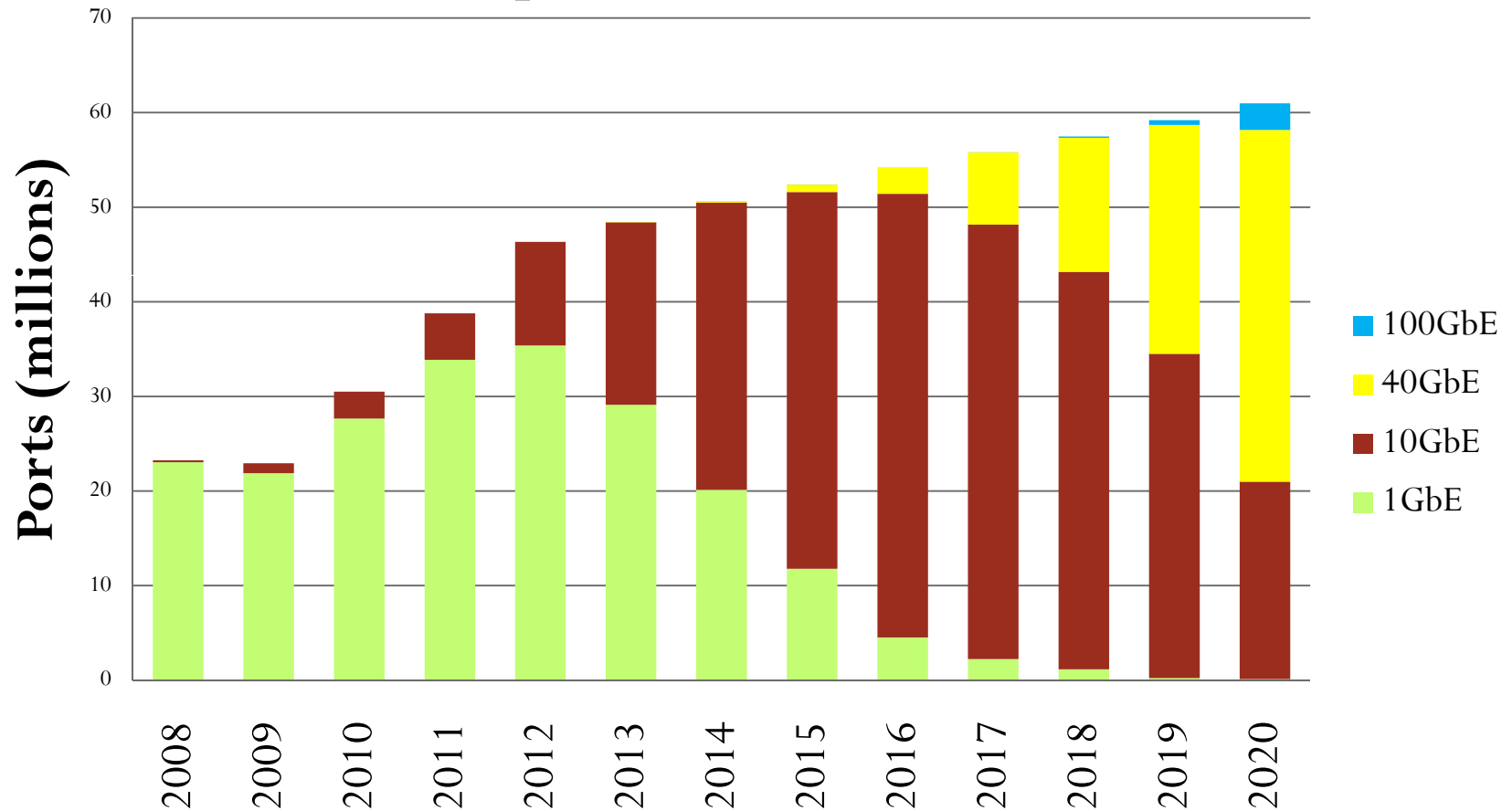…but I can at least ask for a second opinion

Another Very Long Term Estimate for Server Ports
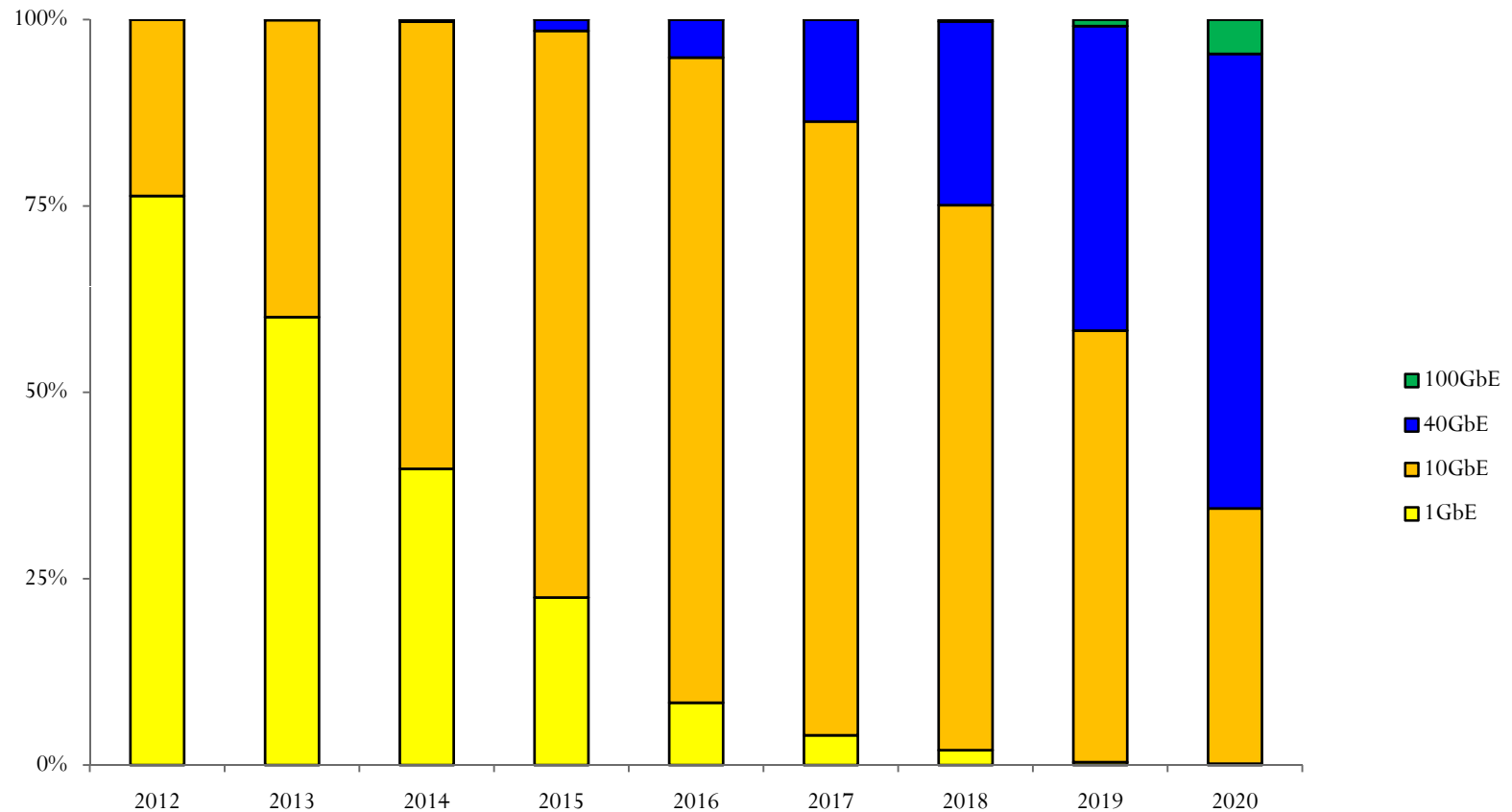
CREHAN RESEARCH Inc.

Legend:
- 100GbE
- 40GbE
- 10GbE
- 1GbE

Server Class Adapter & LOM Ethernet Ports

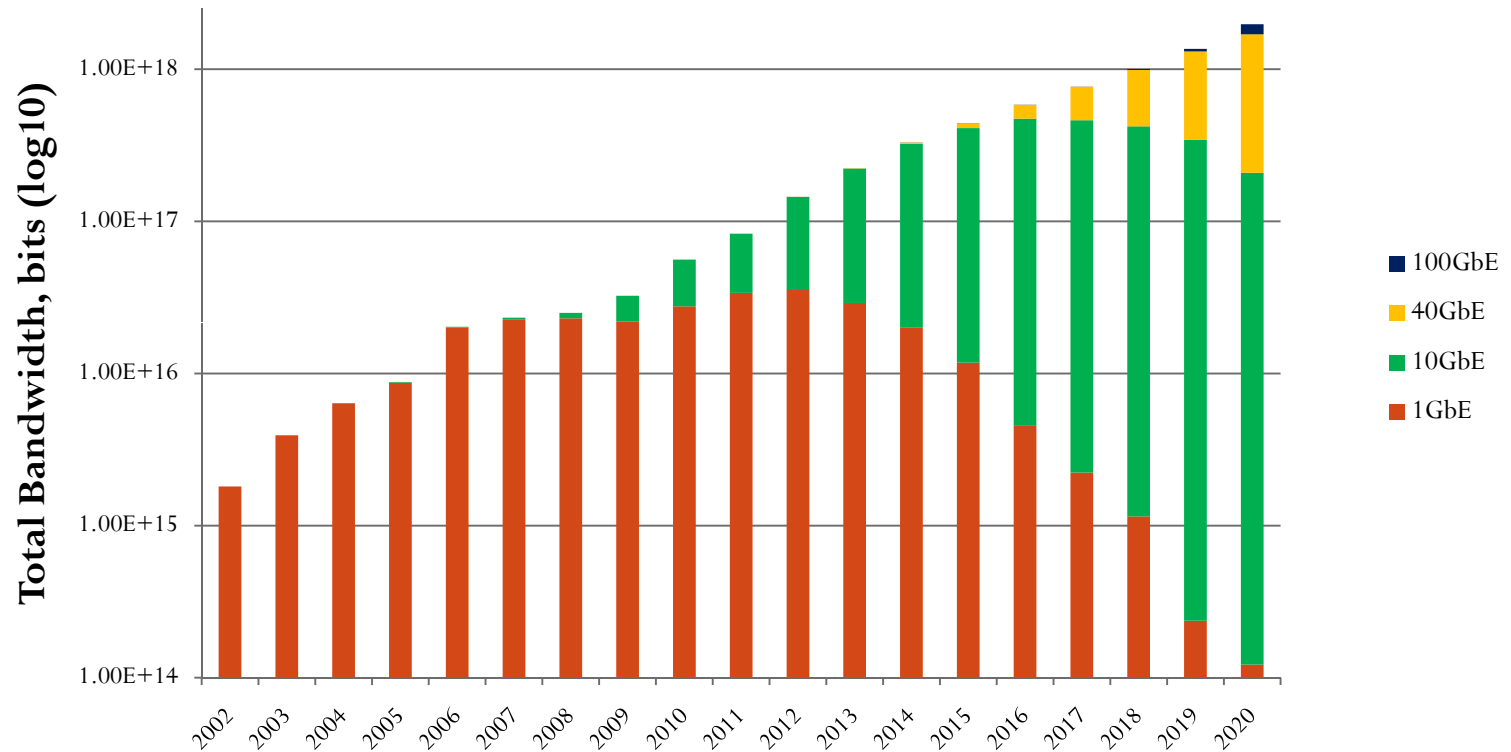Source data: Crehan Research, 2012

What Does That Say About Server Bandwidth Growth?

Server Ethernet Bandwidth Capability

Aggregate port count * port speed

Source: IDC, Dell Oro, Crehan Research.
Reporting methods have changes over the years

Exponential trend continues.

BW doubles every ~2years

Conclusions

Server BW growth continues

- This decade will see 40G & 100G NICs

- Mix is uncertain, but BW need will be there

Next Speed

- 400Gb/s uplinks will serve both 40GbE and 100GbE NICs

- Uplink speed >100Gb/s is *imperative* for 100G NIC usefulness.

  - Needed soon!   …or the guys in .3bj are wasting their time.