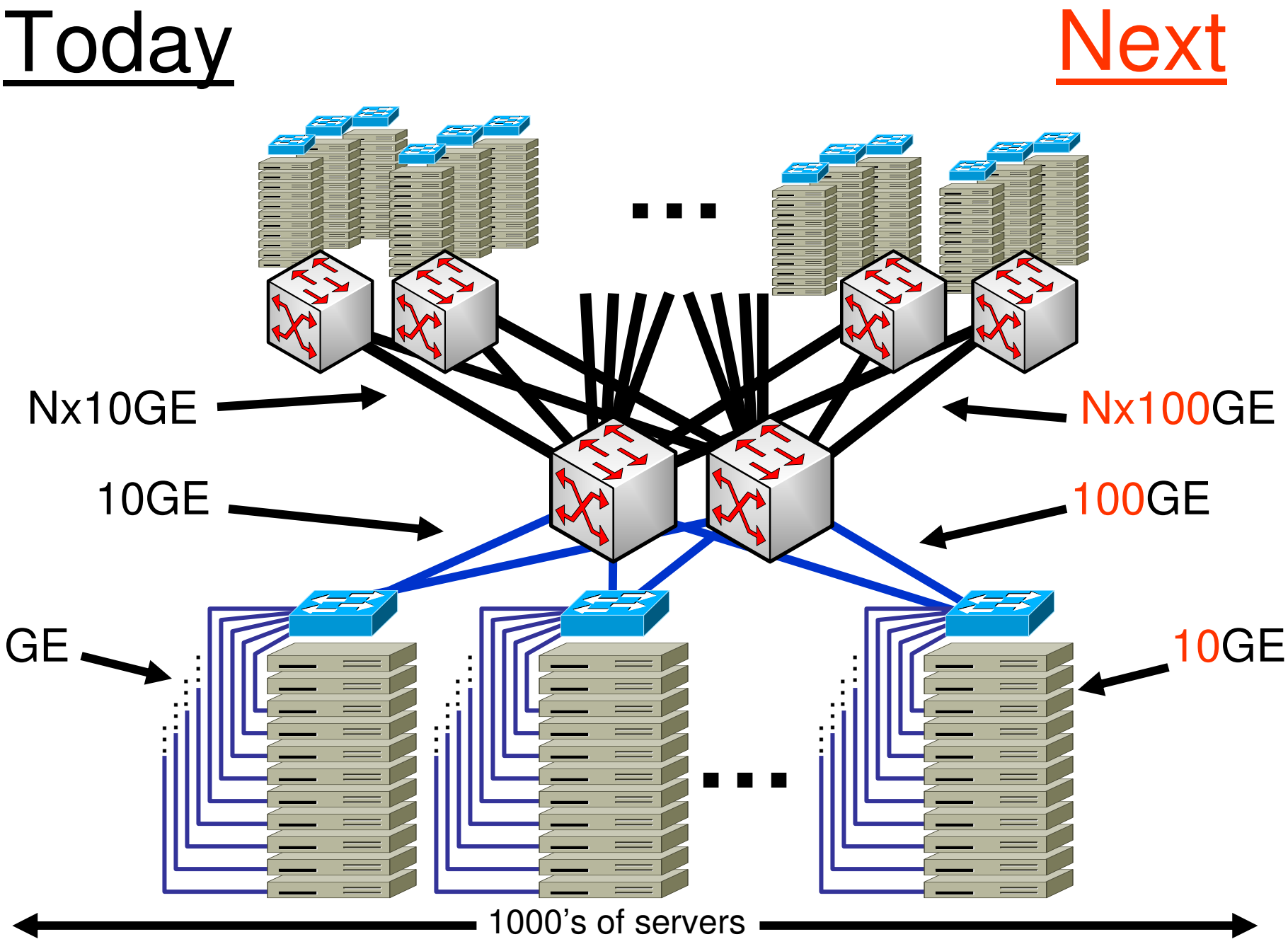
Networking with 1/10/40/100G:
Design and Deployment Perspective

Donn Lee
Network Architecture Team
Google Inc.

My background

- Fortune 50, Consumer Products Conglomerate: Large corporate datacenters, Largest Ungermann-Bass Ethernet deployment
- StorageTek: Large server/storage datacenters, Corporations and Federal customers
- Cisco: Fortune 50 Global customers, Large datacenter/campus Ethernet networks
- Google: Large datacenters, Large networks

Datacenters


Today
Next
Nx10GE
Nx100GE
10GE
100GE
GE
10GE
1000's of servers

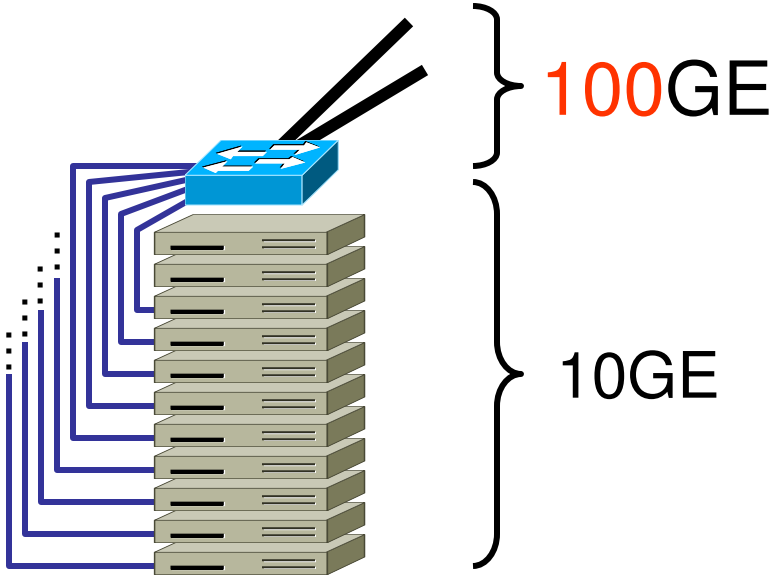
“Amazon is already a significant user of 10GE and Nx10GE. We foresee a need for 100GE soon and encourage the IEEE to begin standardization work immediately."

-- Dr. Werner Vogels

-- Amazon CTO

Hosts are driving 100GE

- GE hosts already driving 100GE aggregation
- 10GE hosts drive more bandwidth
- Cannot aggregate 10GE hosts with 40GE uplink
- 100GE maintains bandwidth hierarchy & oversubscription ratios
- This relegates 40GE to a host-only application



Why embark on a host-only effort when previous standards maintained a general-purpose viewpoint?

Re: Hosts not able to push 100G

- There has never been an expectation that hosts should be able to saturate a nascent LAN standard
- Example, early GE NICs: Hosts could only push 300-400Mbps
- No one faulted the IEEE for overlooking “400M Ethernet” or “4G Ethernet”
- Hosts caught-up in time for user expectations

Just because it can be done, doesn’t mean there’s a big market (or it should be done)

Upgrade path must make sense

- Hosts at 10GE are done, Nx10GE LAG (802.3ad) hosts work fine. What's next is what we care about
- 2-port 10GE NICs available today. Two NICs satisfy the 40G requirement
- 40GE adoption will likely stall: 4x10GE costs less
- Won't be able to LAG 40G since 2x40G=80G when 100GE is shipping
- CIOs do not want two upgrade cycles when they can pay for one

From someone who has many servers

"It takes a number of years to cycle switch infrastructure (and router uplinks) up one speed step. While 40G 'appears' to fill the gap by making the step smaller, in point of fact it makes the step too small, forcing two replacement cycles."

Richard Colella
VP, Network Engineering and Operations
AOL

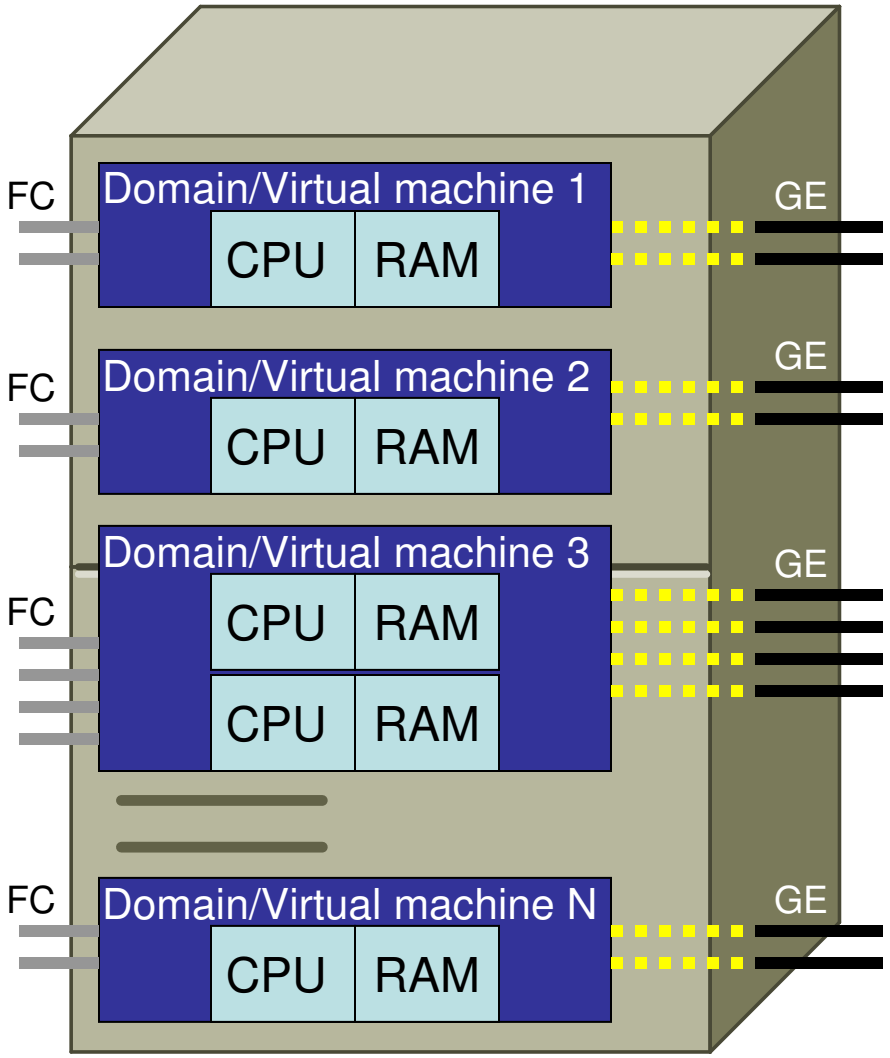
802.3ad LAG was invented to provide a stop-gap measure

- So we don't have to keep inventing incremental PHYs
- If you want a stop-gap measure, the IEEE has provided one
- End-users embrace it
- It drives volumes of GE & 10GE higher
- If you need a 20G, 30G, or 50G server, you can do it
- Please show me the users who say 4x10GE LAG is unacceptable at the host

"We need to focus our energy on 100G."

Vik Saxena, Ph.D.
Senior Director, Network Architecture
Corporate CTO Office
Comcast Cable

What about large servers?



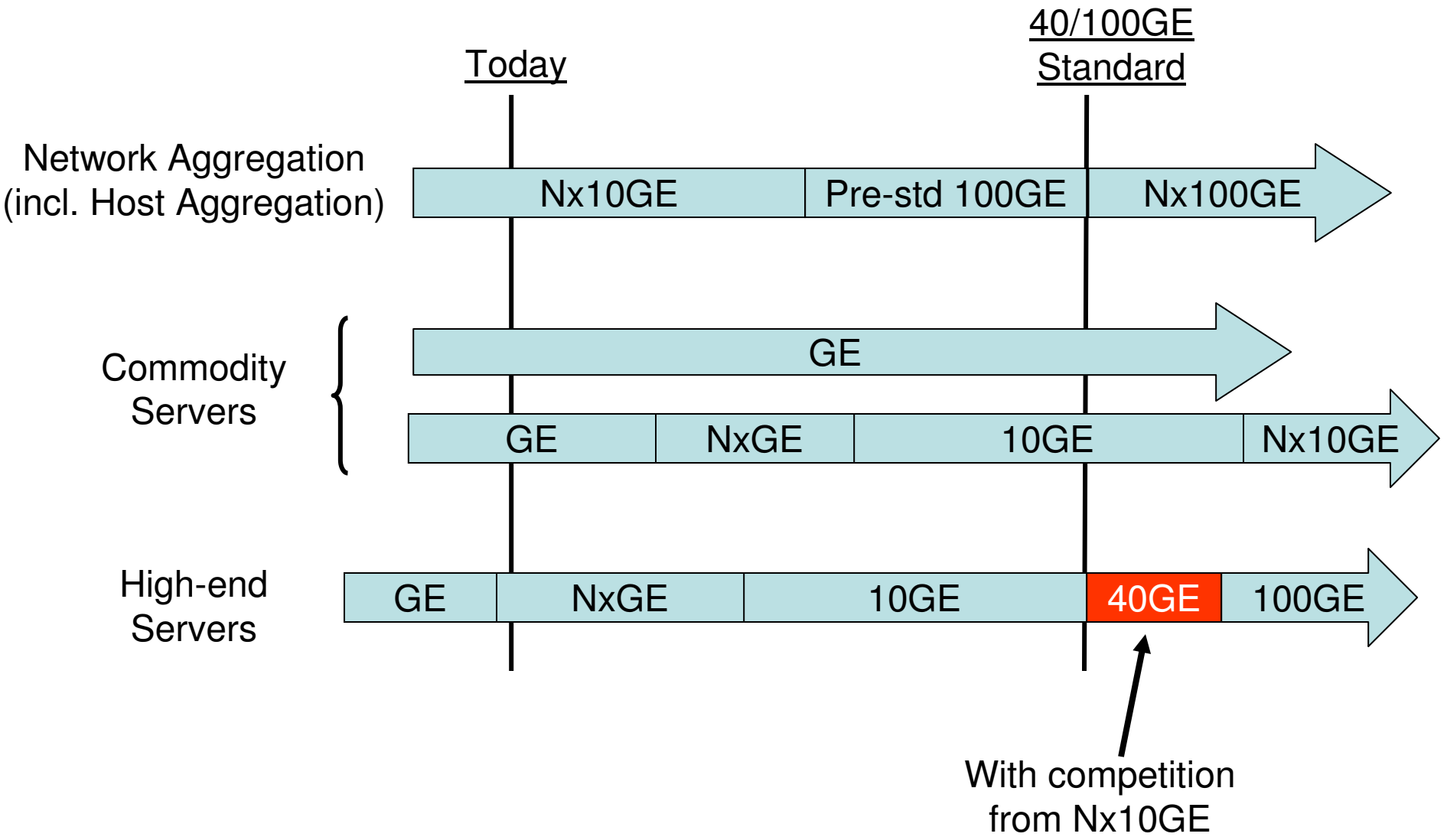
1GE ports assigned to virtual-machines

Strict increments of smaller NICs is preferred over one large pipe (dedicated & secure)

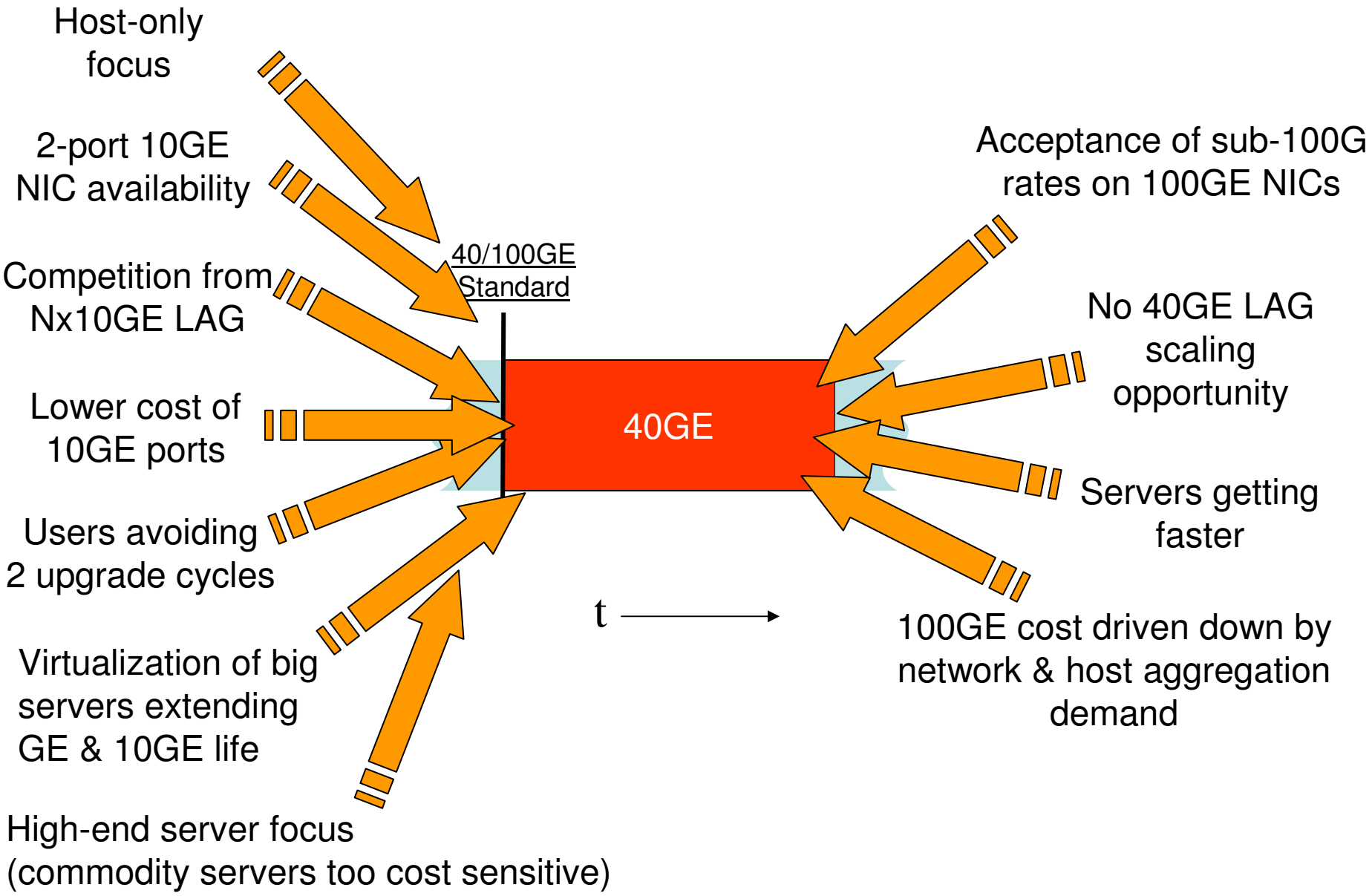
Actual
Sun Fire E25000
Case study

Refer to *lee_01_0307.pdf*, HSSG, 3/9/2007
Re: Virtualization of 1GE driving 100GE urgency

40GE market window



Factors constraining 40GE market window



Summary

- 40GE not useful at the aggregation layer
- Don’t see 40GE advantages over 4x10GE at the server
- Rather have 100GE interface and let the host catch-up
- IEEE 802.3ad LAG provides stop-gap rates between 10G - 100G
- 40GE market window compressed on front-edge and trailing-edge
- Why develop a class of products with narrow focus and limited lifetime?

Thank you

Supporters

- Victor Blake, Advance/Newhouse Communications
- Alan Judge, Amazon.com
- Brad Booth, AMCC
- Henk Steenman, AMS-IX
- Richard Colella, AOL
- Jay Moran, AOL
- Hugh Barrass, Cisco Systems
- Mark Nowell, Cisco Systems
- Vik Saxena, Comcast
- Jason Weil, Cox Communications
- Dan Dove, Dove Networking Solutions, ProCurve Networking by HP
- Greg Hankins, Force10 Networks
- Greg Chesson, Google
- Bob Felderman, Google
- Stephen Stuart, Google
- Drew Perkins, Infinera
- John Jaeger, Infinera
- Larry Green, Ixia
- Michael Bennett, Lawrence Berkeley Laboratory
- Parantap Lahiri, Microsoft/MSN
- Brian Swenson, Microsoft/MSN
- Blaine Christian, Microsoft/MSN
- Christian Nielsen, Microsoft/MSN
- Brent Draney, NERSC
- Peter Harrison, Netflix
- Vish Yelsangikar, Netflix
- Peter Schoenmaker, NTT America
- Andy Bach, NYSE Euronext
- Ted Seely, Sprint
- Bill Trubey, Time Warner Cable
- Adam Bechtel, Yahoo!