Saturating 100G and 1T Pipes

Donn Lee
Network Architecture
Google, Inc.
March 13, 2007

Virtualization: Scaling application resources

Brute force to overcome machine limitations is feasible

Clusters alleviate CPU, storage, and network bottlenecks

- Collective machine bw scales well
- Fiber, lambdas, router ports not as easily

Need for faster servers addressed by massive arrays of commodity machines

Virtualization: Scaling application resources (2)

Machine is no longer the server. The cluster has become the ‘server’

Looking at NIC speed from the old paradigm doesn’t apply to today’s Internet datacenters

Network aggregation layers are the new ‘NICs’ (pathways to the server)

Typical figures on datacenters are an old paradigm for the Internet datacenter

Virtualization: Bottom Line

Machine NIC has not been a serious bottleneck

Possible to get reasonable volume of 10GE based on today's massive aggregation of 1GE machines

10GE port shipments would be even higher if hw vendors could develop mega-sized switches fast enough (demand exists)

Datasets have changed

Size

- Video content x times bigger than its html+image analog
- Satellite imagery databases
- Large files, backups, archives

Machines/services geographically dispersed

Datasets more transient/nomadic, host-agnostic

User expectations always increasing

- Less waiting, Higher resolutions

Aggregation layers are strained

Datacenter demands increasing aggressively

10G LAGs are not adequate

- Often limited to 16x10GE
- Operationally cumbersome
- Clos/Benes fabrics get huge fast at today's 10G density

40G LAGs barely enough

- Not worth switching to a slightly better dead-end

Even at 100GE, LAGs will be limiting

Google™

Metro layer can't be neglected

100GE must be DWDM-friendly

Transport systems should abstract massive λ's (10G or 40G) for 100G client interfaces

Develop higher λ density to scale metro in-step with agg layer

Work actively on 100G serial

Google™

Edge bandwidth under pressure

10G transit ports gated by providers' backhaul upgrades

N x 10G transit router → Much bigger backhaul pipe needed

Japan deploying GE to the home

2006: Internet traffic quadrupled (not counting YT)

Google™

Box we could use

Box we could use today*:

- 25 100GE ports for computing resources
- plus, 600G uplink bw
- 3.1T switch

Box we need soon

- 50 100GE ports
- 5T switch
- And would still require a Clos arrangement

*Based on a conservative design with 4:1 oversub

Google

Reach, cables, n' such

The 'typical' datacenter is an old paradigm for intra-datacenter reach

- 400m-1km needed (2km SR?)

Leverage installed fiber plant

- Retrofitting with ribbon cable is not desirable
- 100-200m not enough reach anyway

Neighboring devices are not co-located intentionally

- Redundancy/diversity
- Can't assume "VSR-ish" applications apply

2009 timing: Will be a very uncomfortable wait

Google™

Bottom Line

Need it today. More so next year.

In 2009, start work on H^2SSG

Make it compatible with DWDM infrastructure

Old ways, comparisons do not apply

- Virtualization, Datasets, Datacenters

5T switch in 2009: Highly desirable