Developments in Internet Infrastructure

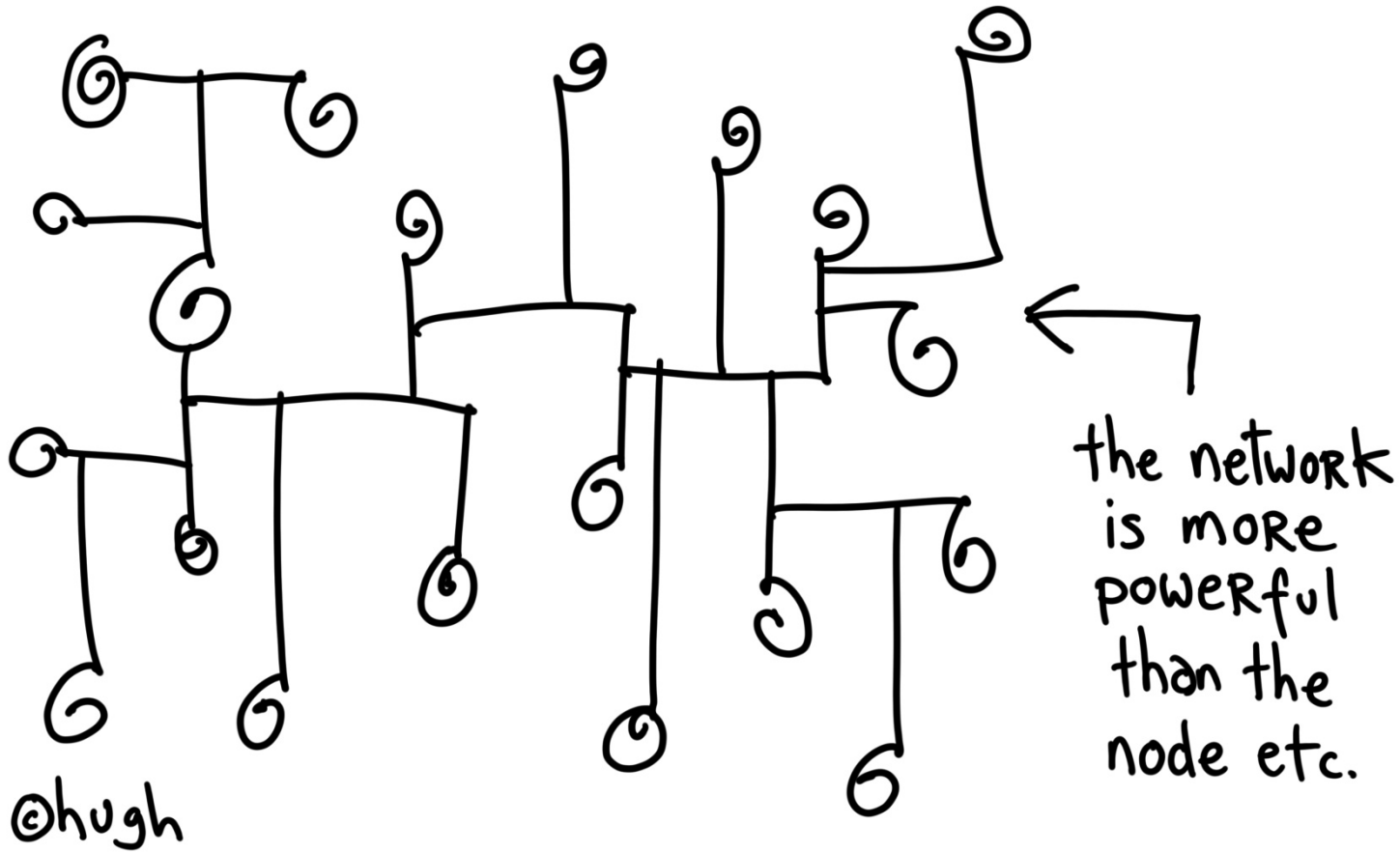
Anne CM Johnson

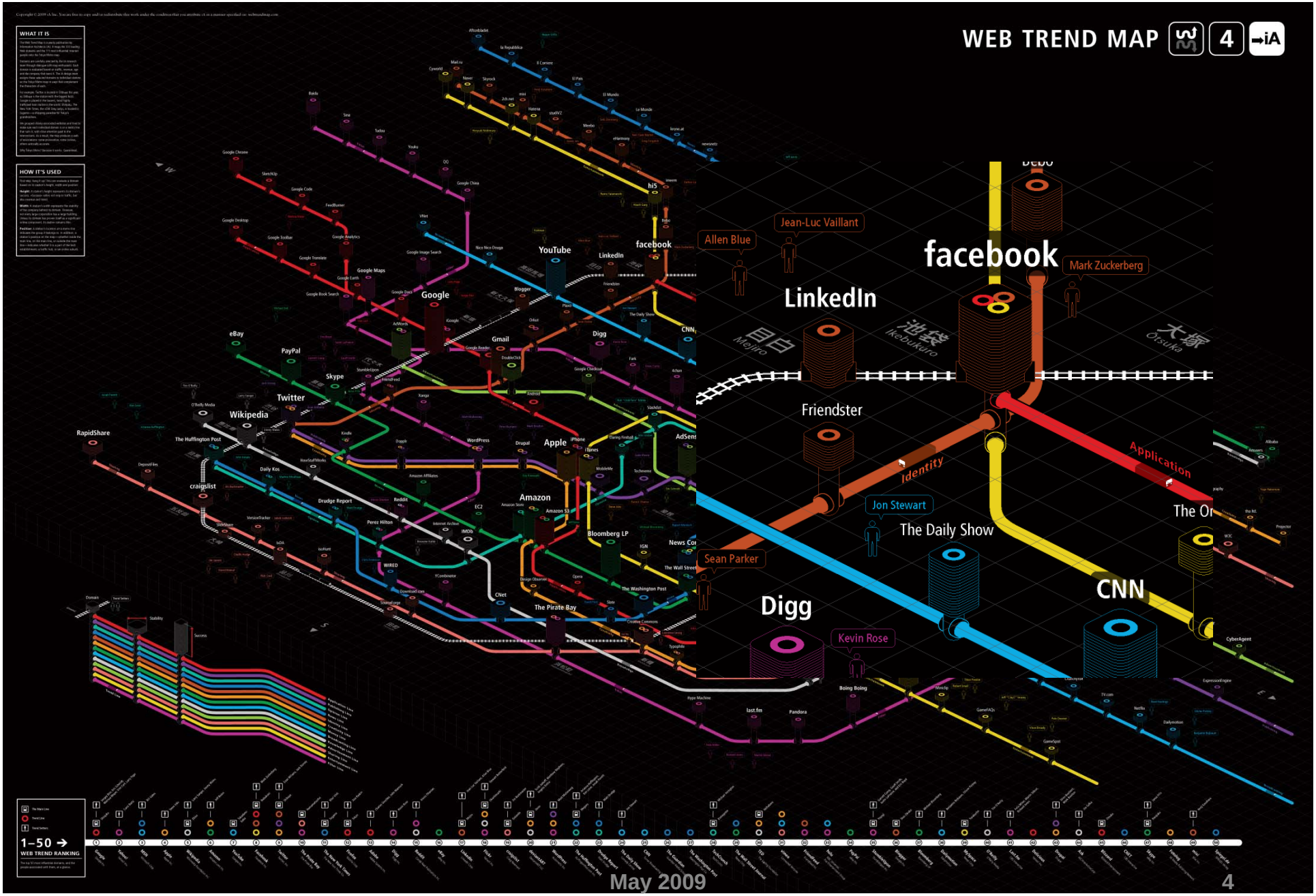
acw@cunningsystems.com, acw@aristanetworks.com, acw@xkl.com

Developments

- Infrastructure
- Latency, bandwidth, diversity
- DWDMs and RAMAN amplifiers
- 10GE and Cloud Computing
- Multi-core, multi-threading
- References

Network Infrastructure

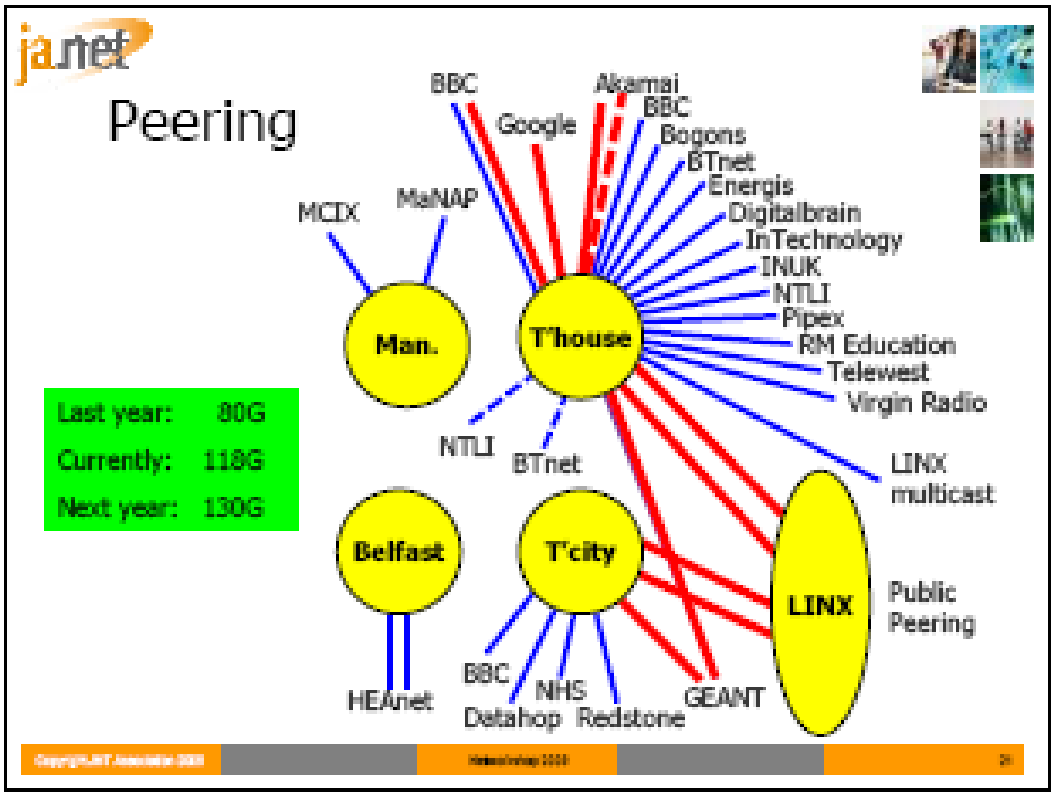



WEB TREND MAP
WHAT IT IS
HOW IT'S USED
facebook
Mark Zuckerberg
Jean-Luc Vaillant
Allen Blue
LinkedIn
Friendster
Identity
Application
Jon Stewart
The Daily Show
Sean Parker
Digg
Kevin Rose
CNN
Boing Boing
Meijiro
Ikebukuro
Otsuka
1–50 →
WEB TREND RANKING
May 2009
4

The Internet /24 IP networks

Image Nov 2003
www.opte.org

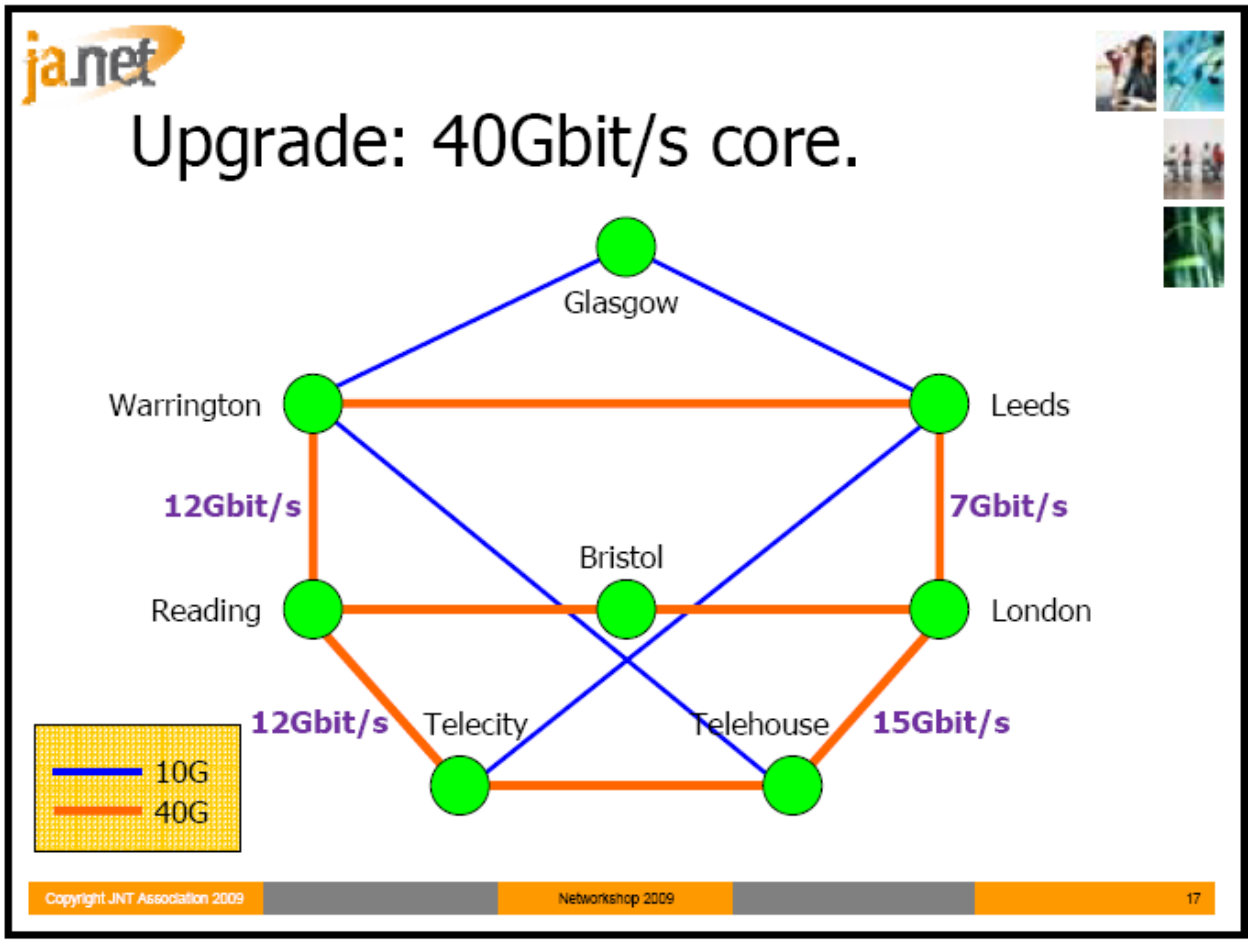
Physical connections - JANET



Peers:
Google
Akamai
BBC

At :
Telehouse
LINX
Telecity

Latest JANET status

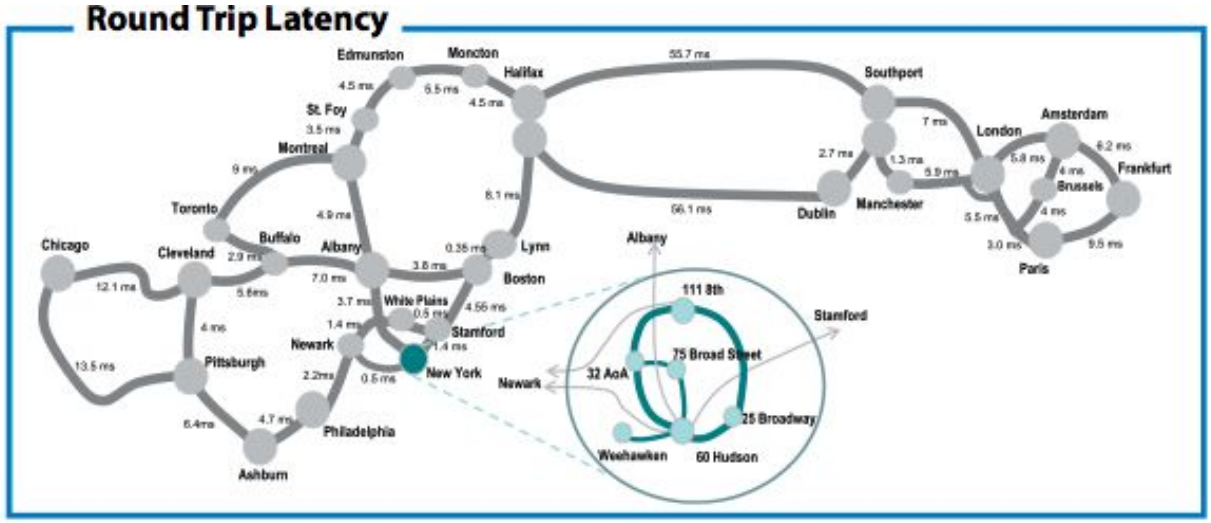


Next upgrade of backbone to 100G

Trials in 2009

Requirement in 2011

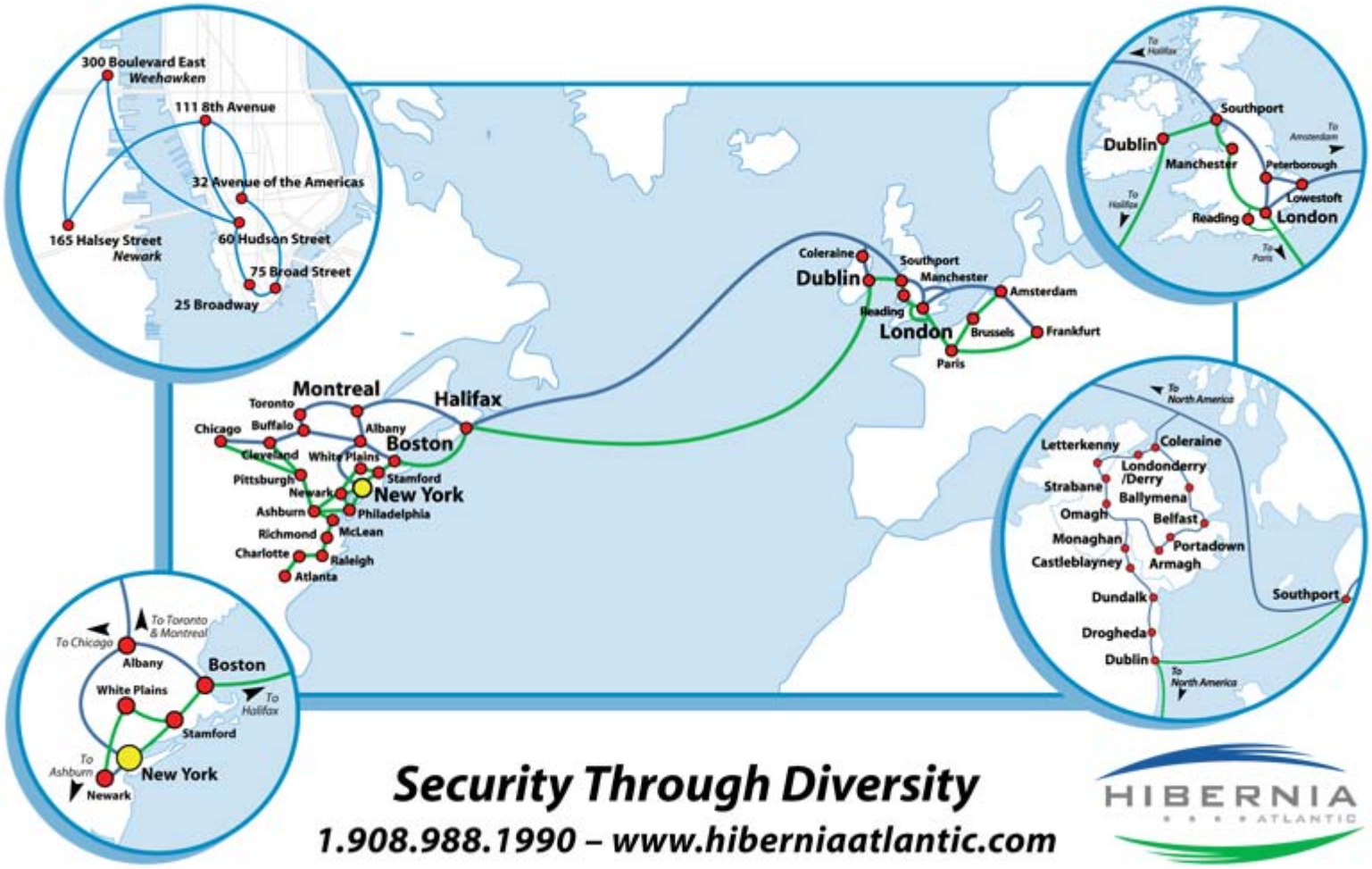
Latency, Bandwidth, Diversity



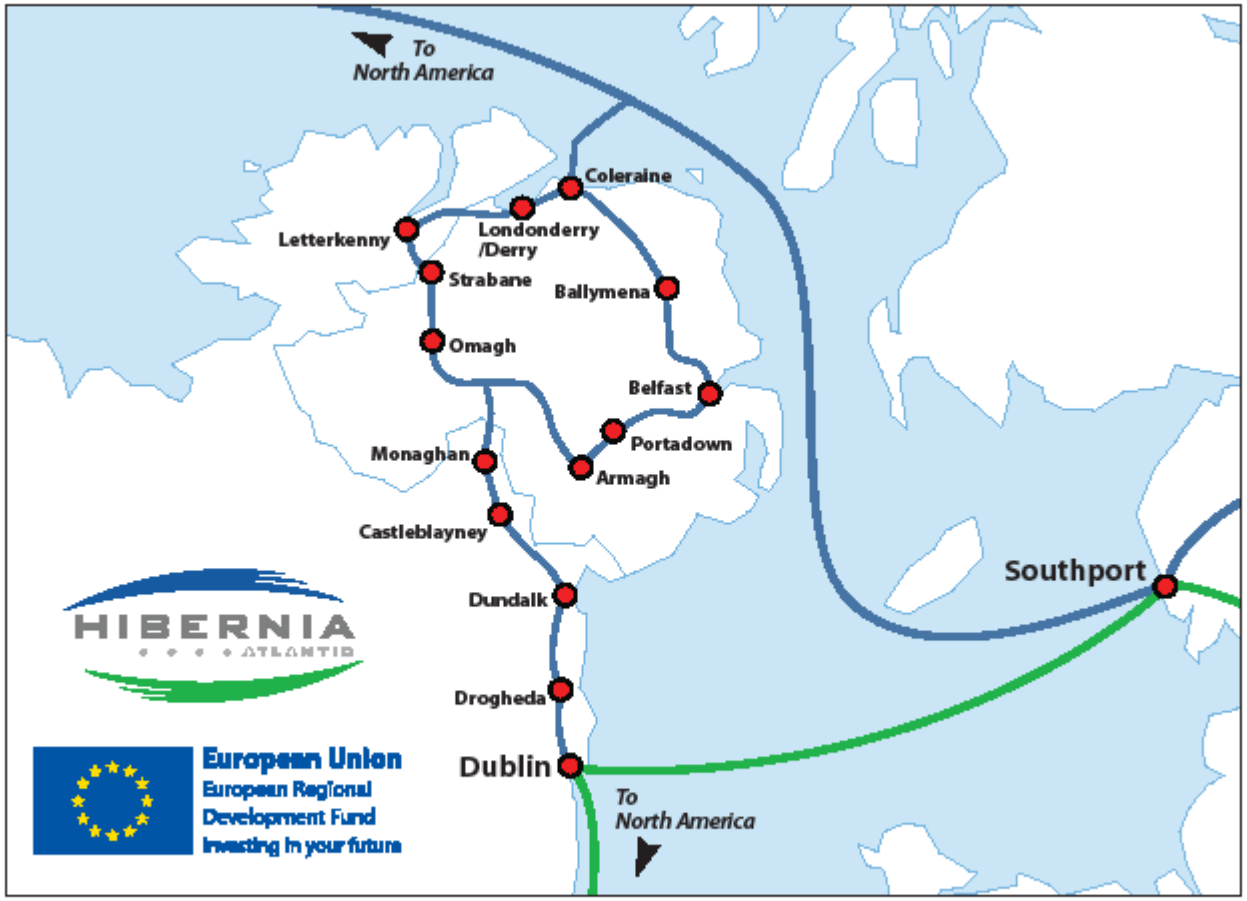
Latency, traffic, attacks

http://www.akamai.com/html/technology/dataviz1.html

Diversity


300 Boulevard East
Weehawken
111 8th Avenue
32 Avenue of the Americas
165 Halsey Street
Newark
60 Hudson Street
75 Broad Street
25 Broadway
Montreal
Toronto
Halifax
Chicago
Buffalo
Albany
Boston
Cleveland
White Plains
Pittsburgh
Stamford
Newark
New York
Ashburn
Philadelphia
Richmond
McLean
Charlotte
Raleigh
Atlanta
Coleraine
Dublin
Southport
Manchester
Amsterdam
Reading
London
Brussels
Frankfurt
Paris
To Halifax
To Amsterdam
To Paris
Peterborough
Lowestoft
To North America
Letterkenny
Londonderry
/Derry
Strabane
Ballymena
Omagh
Belfast
Monaghan
Portadown
Castleblayney
Armagh
Dundalk
Drogheda
To Chicago
To Toronto
& Montreal
To Ashburn
Security Through Diversity
1.908.988.1990 – www.hiberniaatlantic.com
HIBERNIA
ATLANTIC

N Ireland

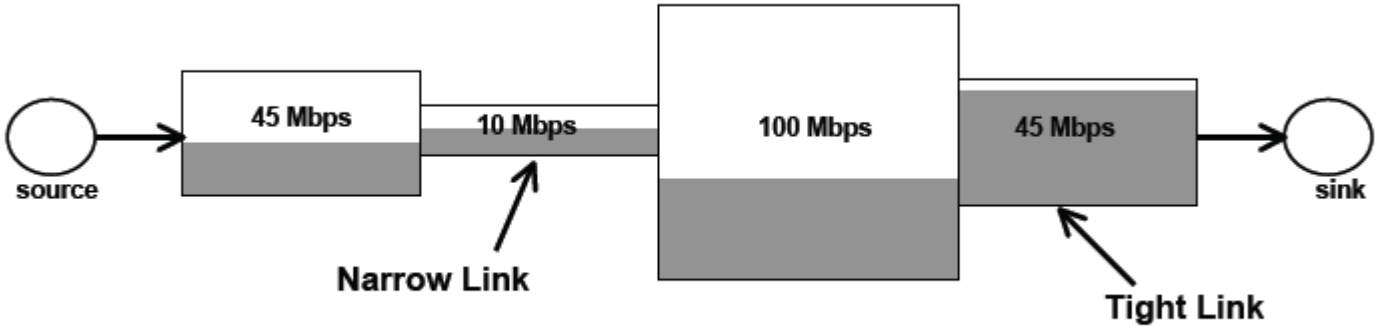


This project is part financed by the European Union's European Regional Development Fund through the INTERREG IVA Cross-border Programme managed by the Special EU Programmes Body.

Fast, reliable network – latency, bandwidth, diversity

Available bandwidth

User experience governed by available bandwidth



Capacity: link speed

- Narrow Link: link with the lowest capacity along a path
- Capacity of the end-to-end path = capacity of the narrow link
- **Utilized bandwidth: current traffic load**
- **Available bandwidth: capacity – utilized bandwidth**
- Tight Link: link with the least available bandwidth in a path
- **Achievable bandwidth: includes protocol and host issues**

http://fasterdata.es.net/talks/Bulk-transfer-tutorial.pdf

wide area bulk data transfers

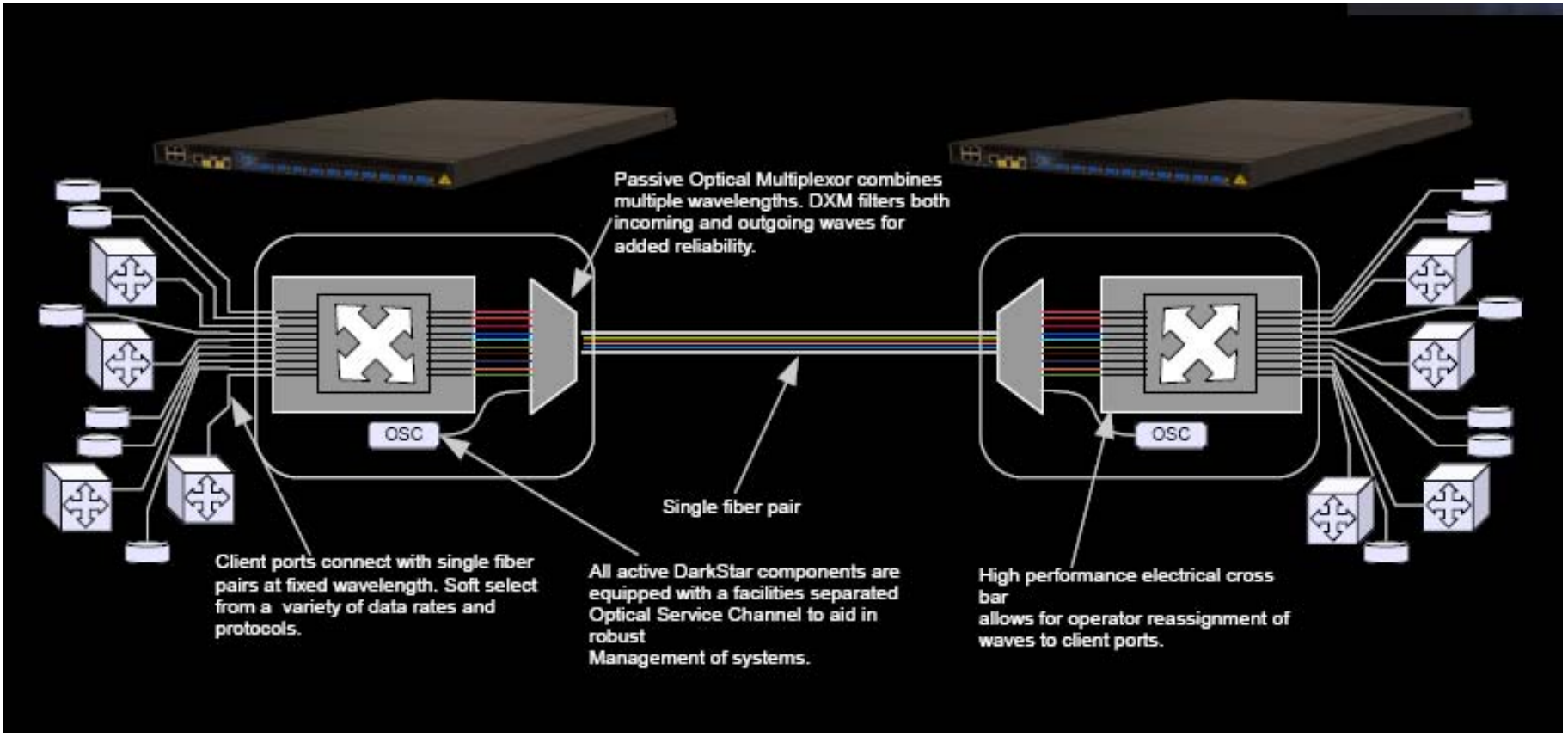
Bandwidth Requrements to move Y Bytes of data in Time X

Bits per Second Requirements

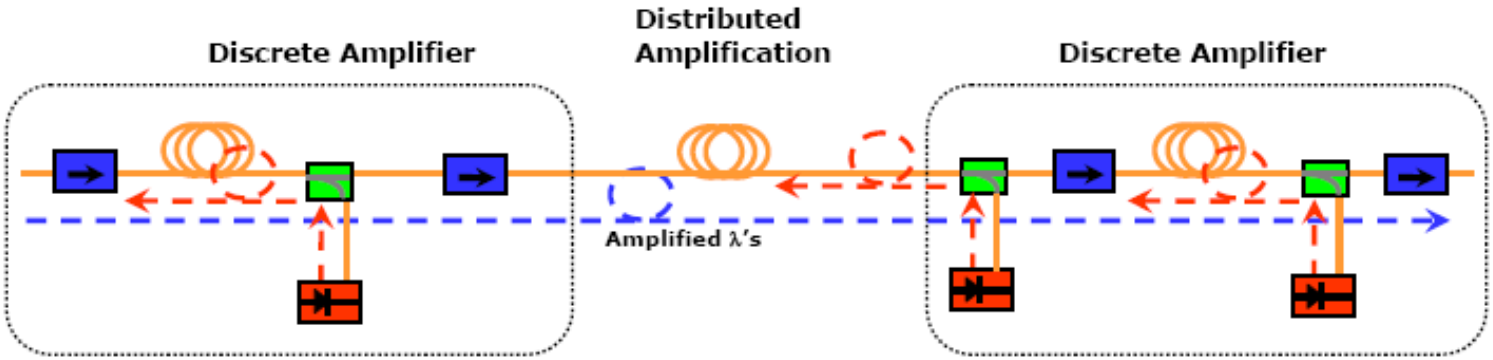
10PB	25,020.0 Gbps	3,127.5 Gbps	1,042.5 Gbps	148.9 Gbps	34.7 Gbps
1PB	2,502.0 Gbps	312.7 Gbps	104.2 Gbps	14.9 Gbps	3.5 Gbps
100TB	244.3 Gbps	30.5 Gbps	10.2 Gbps	1.5 Gbps	339.4 Mbps
10TB	24.4 Gbps	3.1 Gbps	1.0 Gbps	145.4 Mbps	33.9 Mbps
1TB	2.4 Gbps	305.4 Mbps	101.8 Mbps	14.5 Mbps	3.4 Mbps
100GB	238.6 Mbps	29.8 Mbps	9.9 Mbps	1.4 Mbps	331.4 Kbps
10GB	23.9 Mbps	3.0 Mbps	994.2 Kbps	142.0 Kbps	33.1 Kbps
1GB	2.4 Mbps	298.3 Kbps	99.4 Kbps	14.2 Kbps	3.3 Kbps
100MB	233.0 Kbps	29.1 Kbps	9.7 Kbps	1.4 Kbps	0.3 Kbps
	1H	**8H**	**24H**	**7Days**	**30Days**

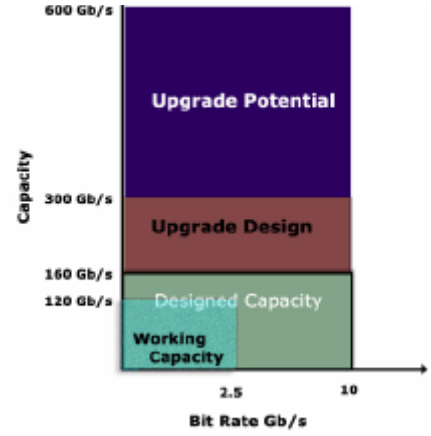
DWDM - XKL

Dense wave division multiplexor



Raman Amplification





Upgrade unrepeated and short repeated links

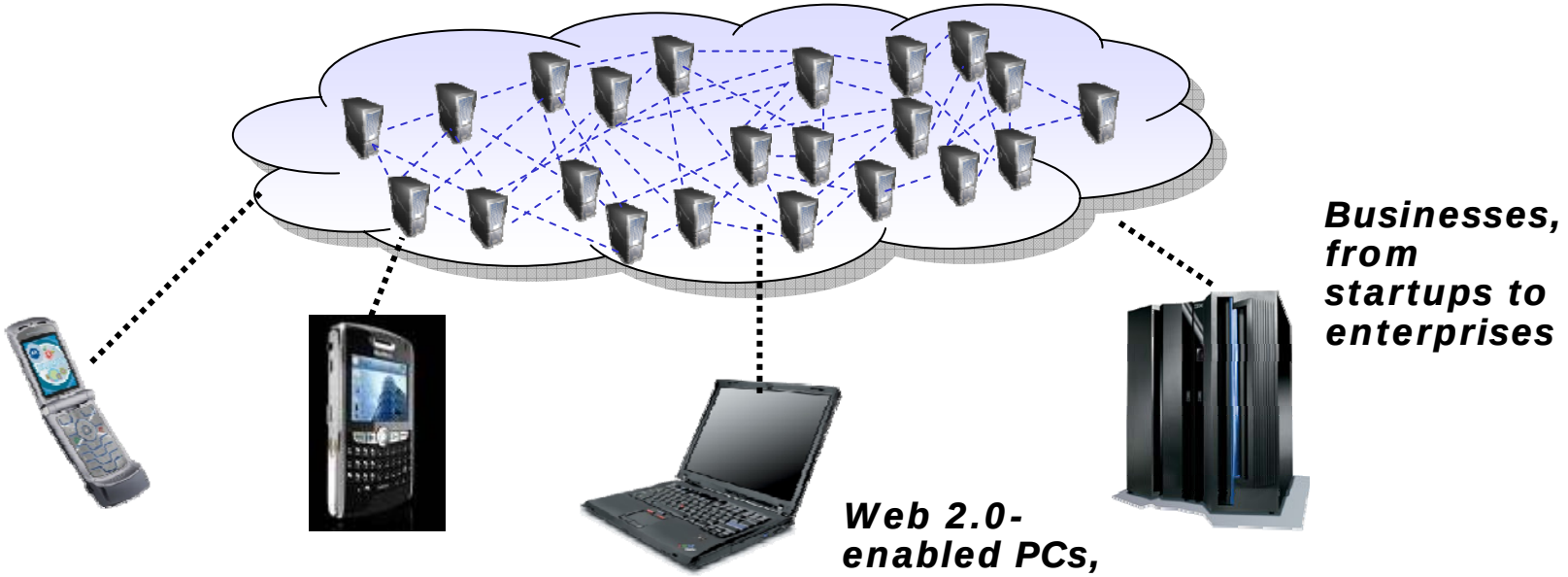
Credit XTERA

Clouds, clusters, containers

- Cloud computing
- Server interconnect
- Building data centers

What is “Cloud *Computing*”?

An emerging computing paradigm where data and services reside in massively scalable data centers and can be ubiquitously accessed from any connected devices over the internet.



Businesses, from startups to enterprises

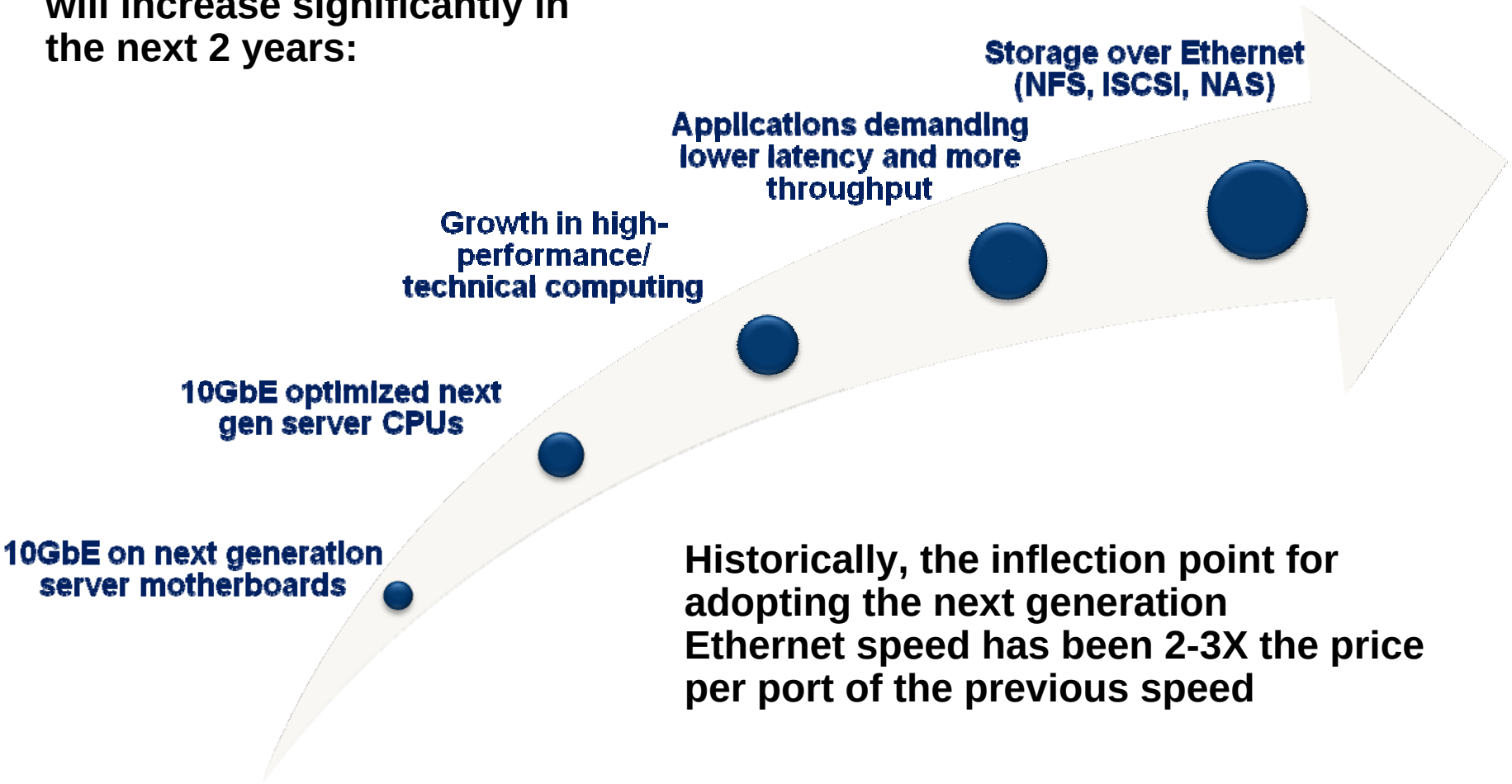
4+ billion phones by 2010

Web 2.0-enabled PCs, TVs, etc.

Source: IBM

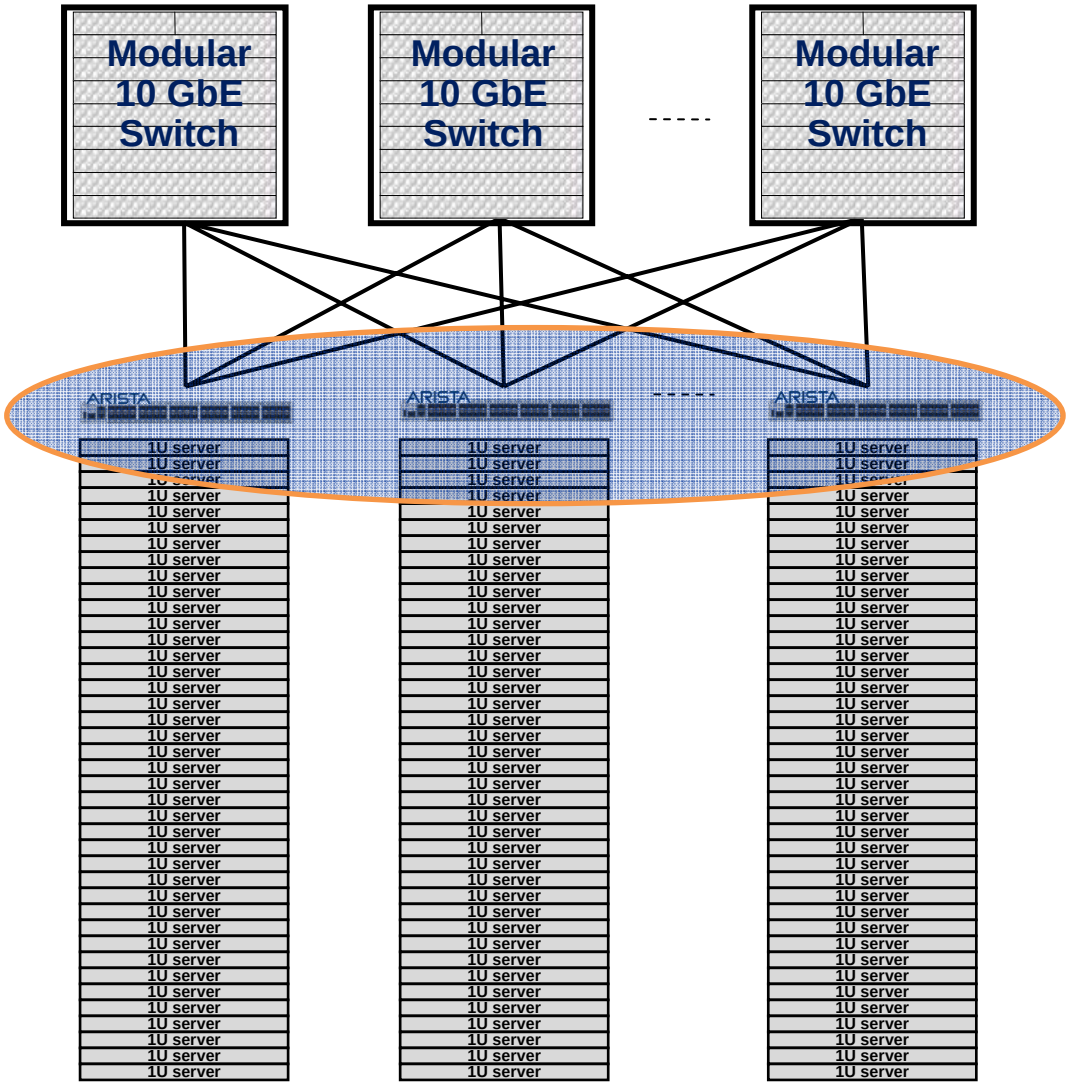
10 GbE Market Growth Drivers

10 GbE server attachment will increase significantly in the next 2 years:



Historically, the inflection point for adopting the next generation Ethernet speed has been 2-3X the price per port of the previous speed

Rack-Server Aggregation - Arista



Rack-Top Switch 24-48 Ports

- Typically 20-40 servers per rack
- 4 to 8 uplinks
- Copper cables in rack
- Optics to core switches

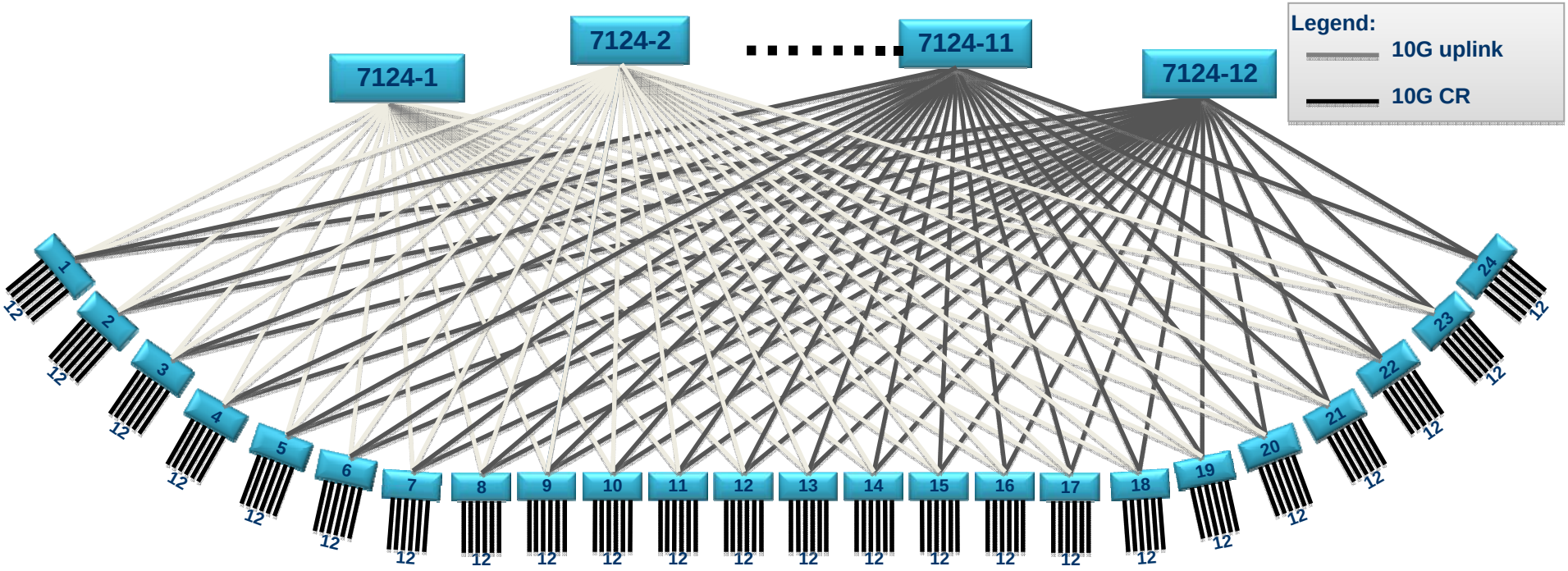
High Availability

- Redundant power/cooling
- Load sharing redundant topology
- Reliable switch OS

Scalability

- High Throughput
- Low Latency

288 port non blocking 10GE cluster



- **12 spine switches, 24 leaf switches, 12 x 24 = 288 10GE ports**
- **Each leaf switch connected to each spine switch via 2X 10G trunk connection**
- **L3 ECMP for multi-path and load balancing**

Data Centers in Containers

Google containers

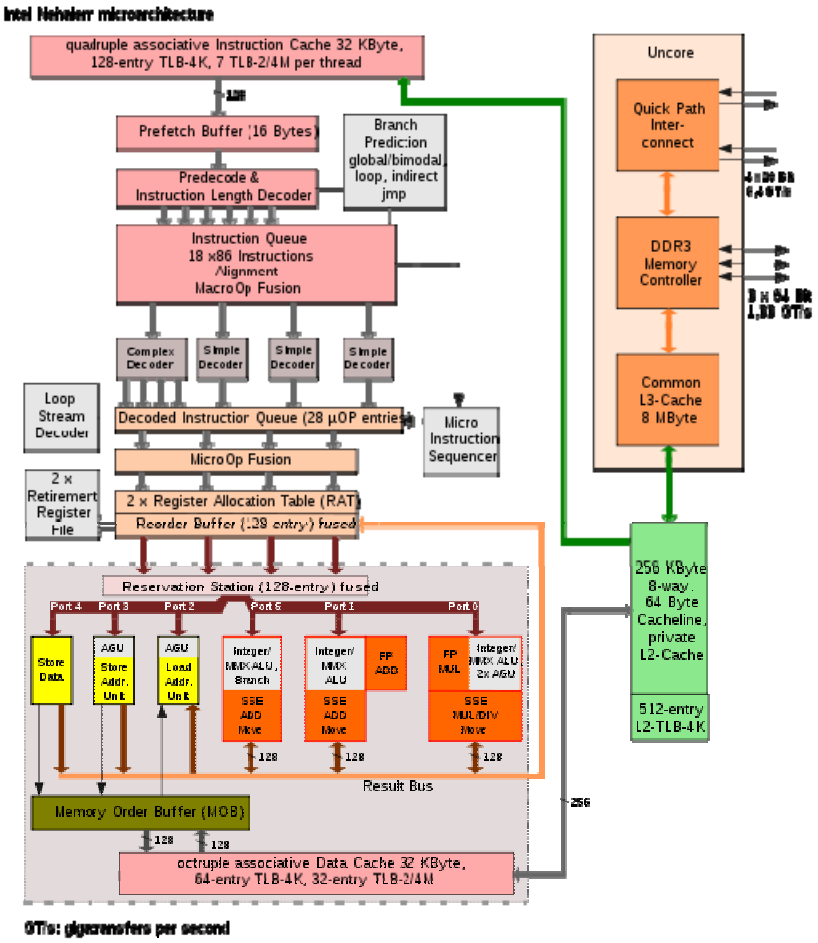
- First used 2005
- Battery backup for each server
- Data center use 1AAA containers (12m) with1,160 servers each
- A facility may have hundreds of containers
- Minimise costs of software, hardware, facilities – revenue per query is low

Inside – HP container

Inside a server

- Started with long haul and UK Wide Area Networks
- Reviewed containers and clusters
- Look at CPU inside the latest servers

Nehalem – Intel Xeon 5500



http://upload.wikimedia.org/wikipedia/commons/6/64/Intel_Nehalem_arch.svg

- Lithography Process: 45 nm
- Cores: 4
- Threads: 8
- Frequency: 2.66 GHz
- Cache: 256 KB L2/core and 8 MB shared L3
- Memory Controller: Triple channel DDR3 800/1066/1333 MHz
- Bus Interface: 1x 4.8 GT/s QuickPath
- TDP: 130W
- Socket: LGA1366
- £200.49 ex vat

Quad core CPU

- Copy of L2 cache on L3 cache
- Memory controller on CPU die not on motherboard – new mechanism (QuickPath) for access to external memory, PCI Express interconnect
- Multi-threading – 2 threads per core
- Changes available bandwidth for application

Response time - server

- Multiple cores, multi-threading support can reduce latency and improve bandwidth
- Serialization has performance implications
- Job service time variation matters too
- Communication bandwidth to cloud
- Where is the bandwidth bottleneck ?

Accelerating old applications

- Motherboards with Nehalem and LOM 10GE
- Existing performance limited programs retesting
- Finite element analysis – automotive crash, powertrain simulation; seismic analysis for oil and gas exploration
- Image processing, video storage and delivery
- Financial automated trading

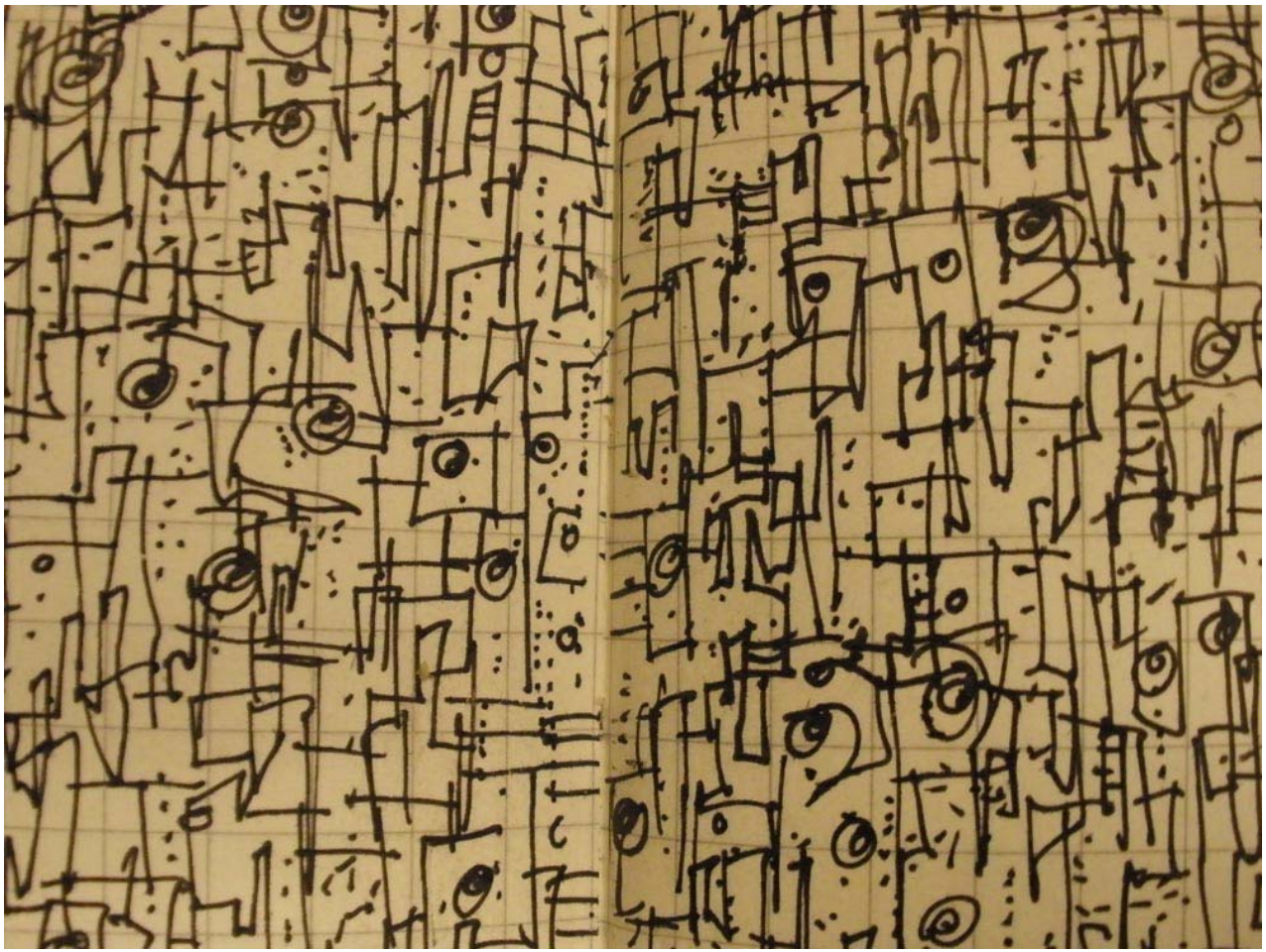
Enabling new applications

- Lifestreams – subscribing to people (David Gelerenter, Mirror Worlds; Robert Scoble, Friendfeed)
- Computing and networking moves ‘from object to fabric’ (Clay Shirky)
- Blogrollr – collects blog reading activity

Summary

- Internet infrastructure is pervasive
- More reliability, bandwidth, less latency improves user experiences
- Huge quantities of processing, bandwidth, storage, memory are available at rapidly reducing cost
- Infrastructure has not yet become ‘fabric’
- Expect more new features and functions

Network System complexity



@hugh MacLeod

References

- http://www.bmc.com/offers/performance/whitepapers/docs/2005/Bandwidth_and_Latency_Their_Changing_Impact_on_Performance.pdf
- http://www.akamai.com/html/technology/dataviz1.html
- Google server designs http://news.cnet.com/8301-1001_3-1020958092.html?tag=newsLeadStoriesArea.1
- Intel Core i7 http://www.purelypc.co.uk/product_detail.php?product=8360
- http://en.wikipedia.org/wiki/Nehalem_(microarchitecture)
- http://www.hpcwire.com/offthewire/Argonne-Speeds-Analysis-Using-MathWorks-Tools-43429957.html
- http://nanog.org/meetings/nanog45/presentations/Tuesday/Chaires_submarine_N45.pdf
- http://mvdirona.com/jrh/TalksAndPapers/JamesHamilton_SMDB2009.pdf
- http://www.edge.org/3rd_culture/gelernter09/gelernter09_index.html