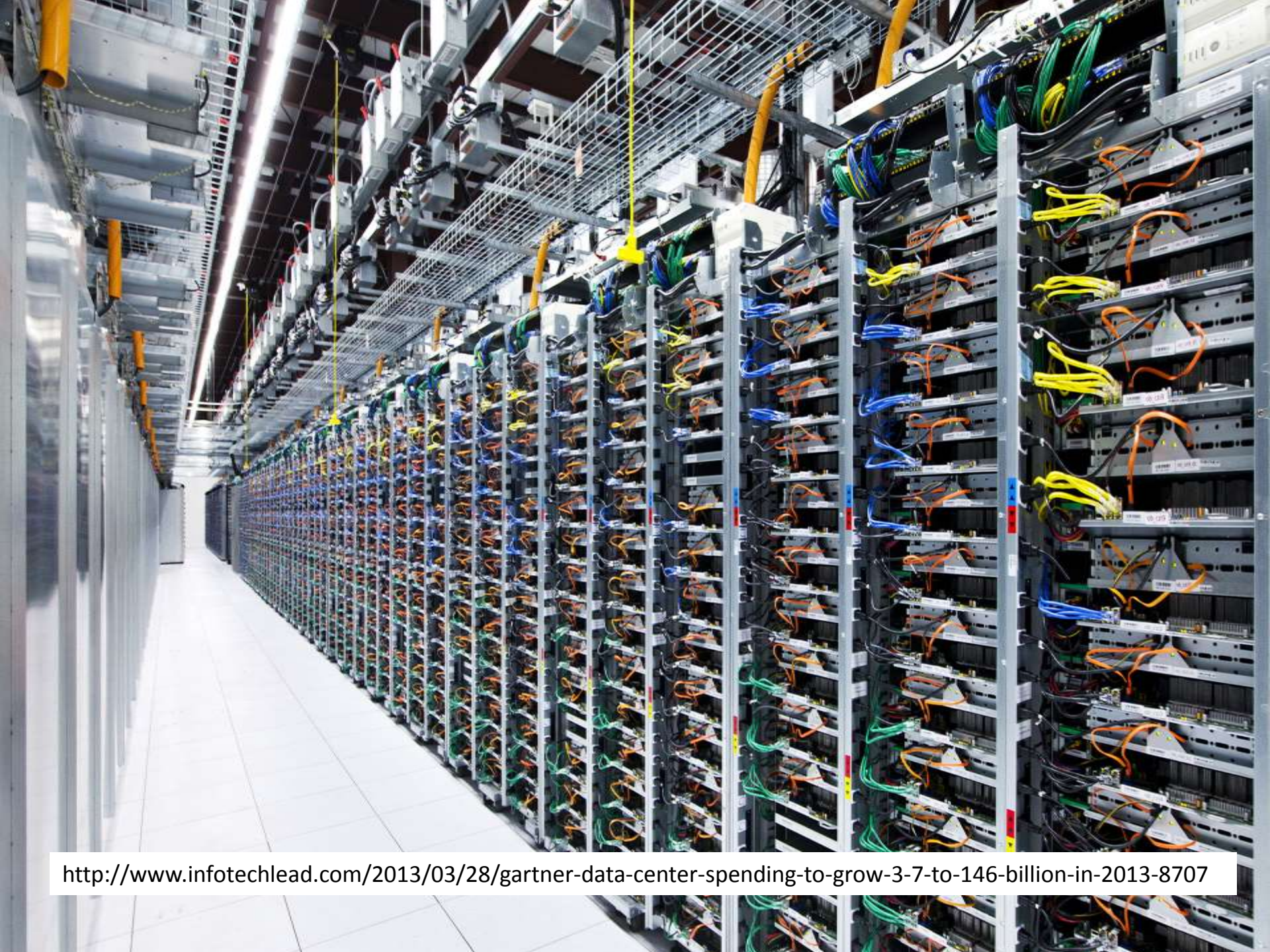


Designing and Experimenting with Data Center Architectures

Aditya Akella
UW-Madison



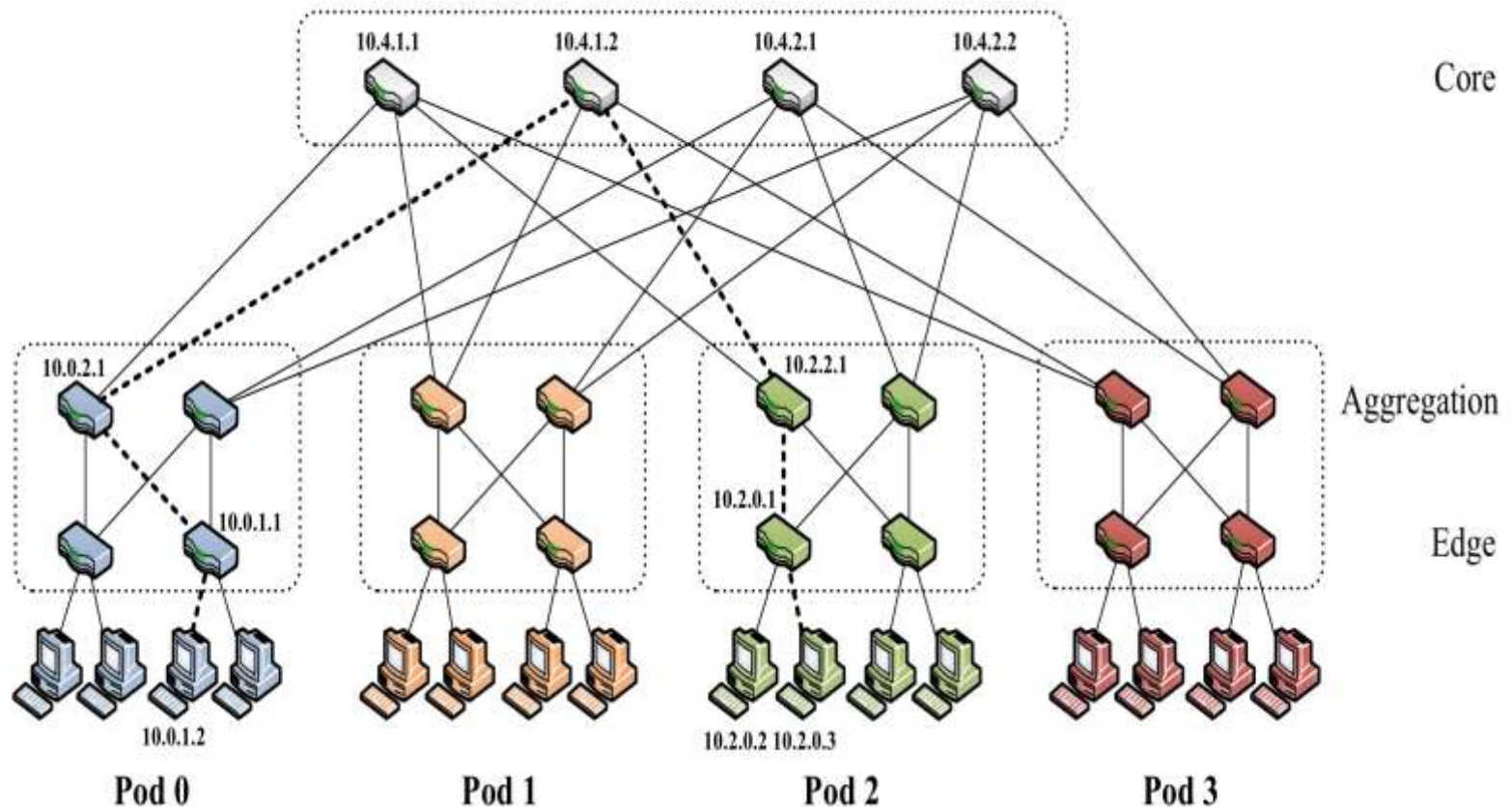
<http://www.infotechlead.com/2013/03/28/gartner-data-center-spending-to-grow-3-7-to-146-billion-in-2013-8707>

What to build?

This question has spawned a cottage industry in the computer networking research community.

- “Fat-tree” [SIGCOMM 2008]
- VL2 [SIGCOMM 2009, CoNEXT 2013]
- DCell [SIGCOMM 2008]
- BCube [SIGCOMM 2009]
- Jellyfish [NSDI 2012]

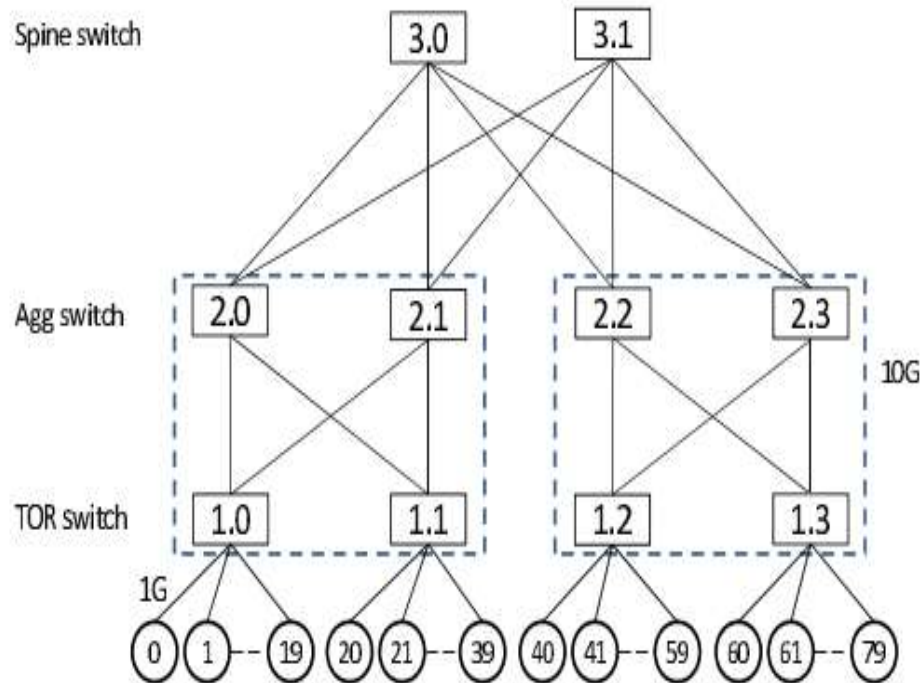
“Fat-tree” SIGCOMM 2008



Isomorphic to butterfly network except at top level

Bisection width $n/2$, oversubscription ratio 1

VL2 (SIGCOMM 2009, CoNEXT 2013)



(b) VL2

called a Clos network
oversubscription ratio 1
(but 1Gbps links at leaves, 10Gbps elsewhere)

How to compare networks?

- Bisection width
- Diameter
- Maximum degree
- Degree sequence
- Area or volume
- Fault tolerance
- Cost

A Universal Approach

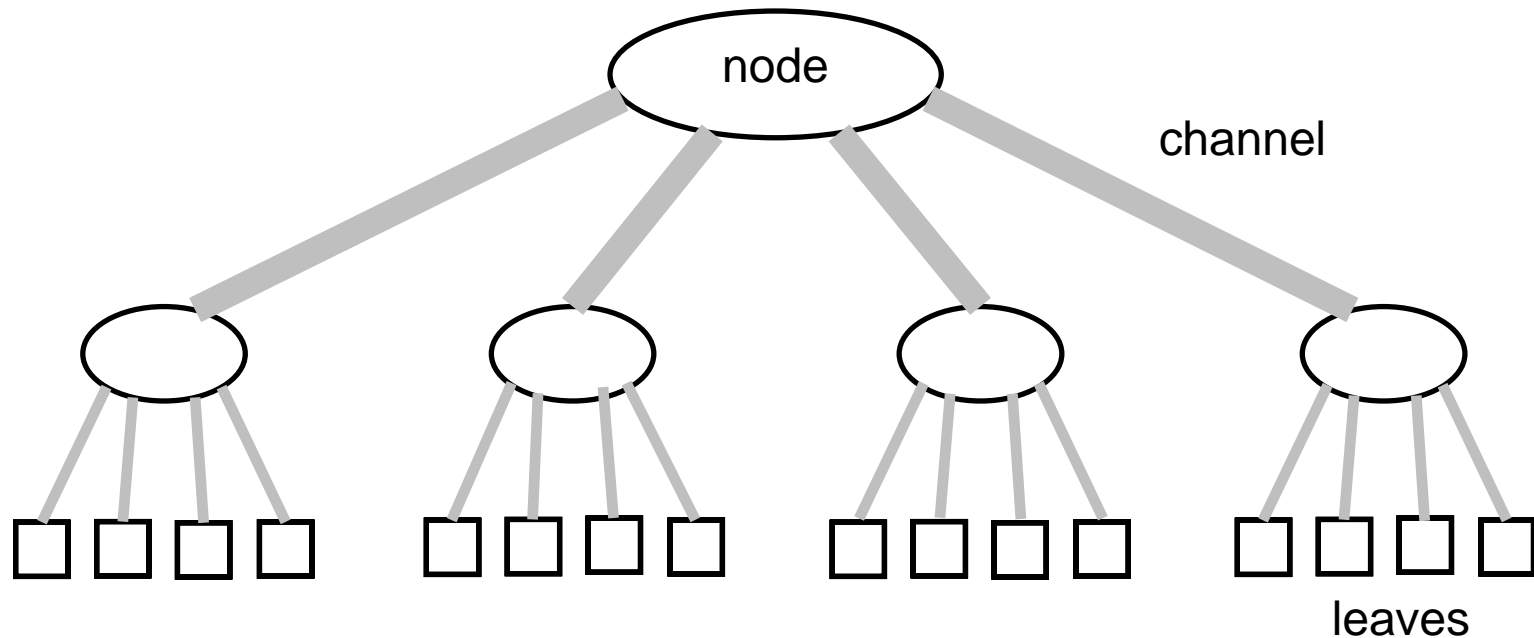
Build a single network that is competitive, for any application, with any other network that can be built at the same cost.

Area-Universality

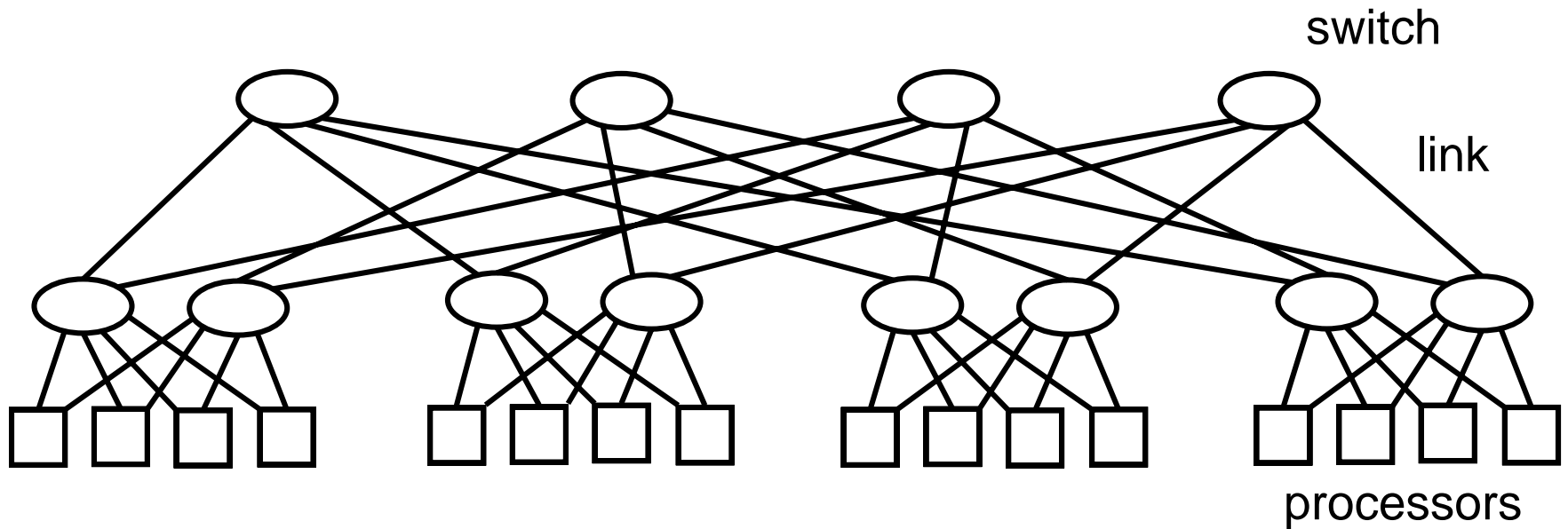
Theorem [Leiserson, 1985]: There is a fat-tree network of area n that can emulate any other network that can be laid out in area n with slowdown $O(\log^3 n)$.

- Later improved to $O(\log n)$ slowdown
- “area” can be replaced by “volume”

Coarse Structure of Fat-Tree Network

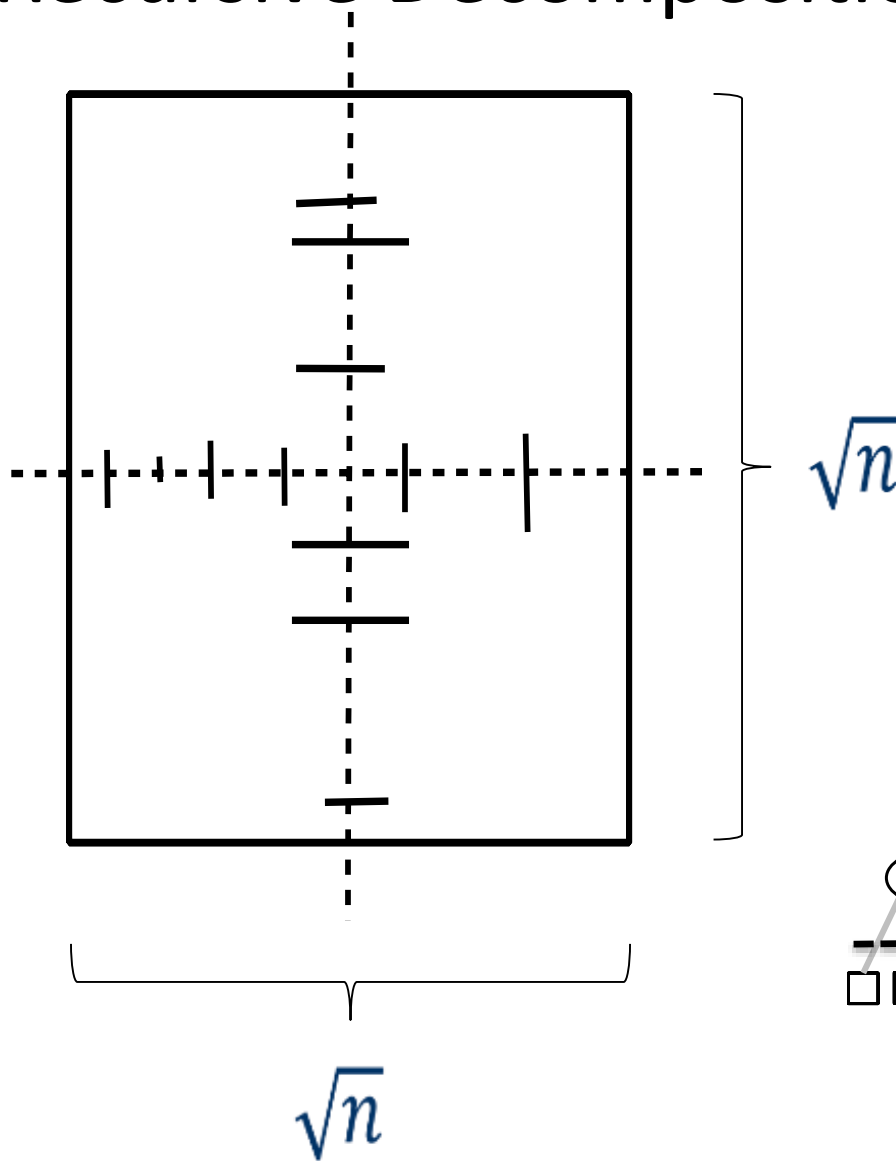


Example of Fine Structure of a Fat-Tree

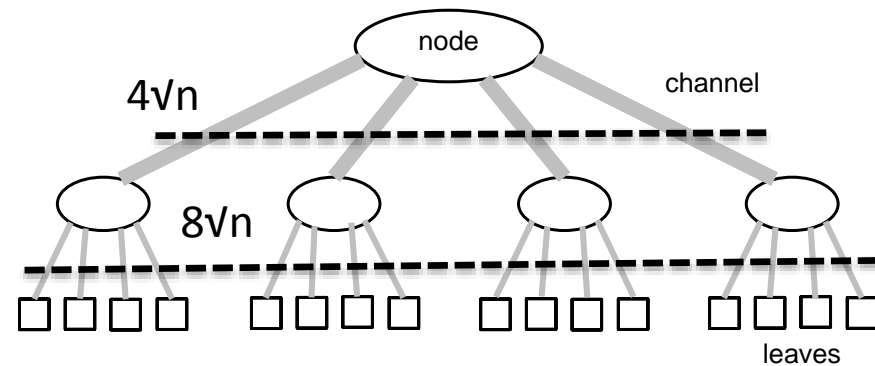


Butterfly Fat-Tree (Greenberg-Leiserson)

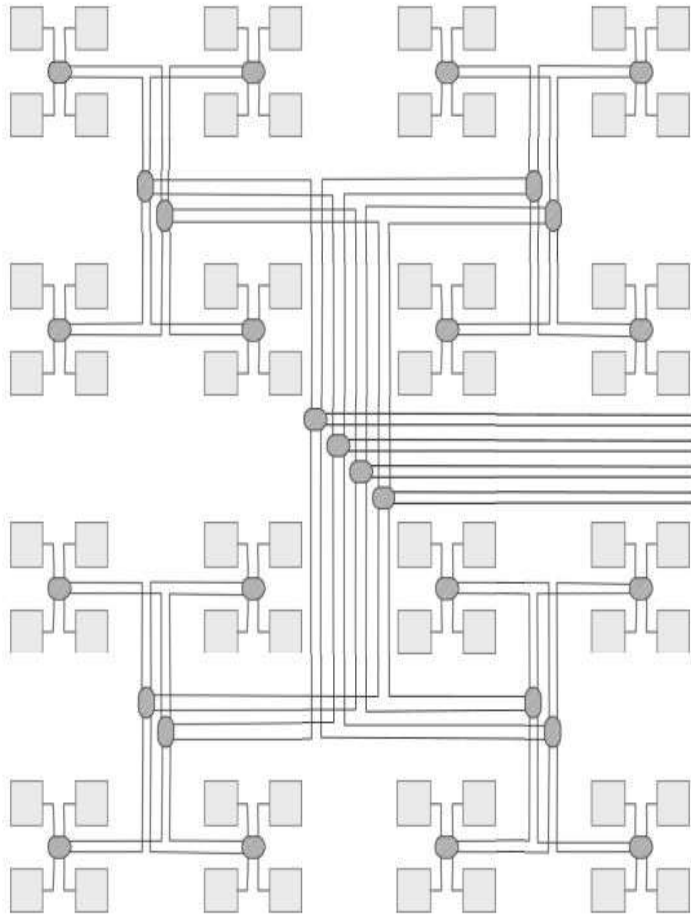
Recursive Decomposition of VLSI Layout



Idea: match fat-tree channel capacity with maximum number of wires cut in layout.



Layout of Area-Universal Fat-Tree



Top layer capacity: $4\sqrt{n}$
Next later capacity: $8\sqrt{n}$
but at half wire length



Equal area at each layer!

Message crossing link in emulated network →
fat-tree sends message across a pair of leaves

The Universal Approach for Data Center Design

Allocate the same amount of **money** to each level in a three- or four-level fat-tree network.

Levels: within a rack, between racks in a row, between rows, etc.

Rule of thumb: build “best” network at lowest level; match cost at higher levels

Caveats

- Assumes that performance is proportional to cost (e.g., for one-third the cost, can buy one-third the capacity)
- Assumes that it is physically possible to spend the same amount at each level
- Assumes that bandwidth bottlenecks, not latency, limit performance.

Trends

Looked at large-scale network deployed by a major provider of on-line services.

Five years ago, money spent on layers (bottom up) was 6:2:1.

In 2014, the ratio was 2:2:1 → we're building nearly cost-universal networks today!

CloudLab



The Need Addressed by CloudLab

- How to optimize software (e.g., to take superfluous latency out)?
- How best to design and run applications?
- Storage, networking, virtualization, ...
- To investigate these questions, we need:
 - Flexible, scalable **scientific infrastructure**
 - That enables exploration of **fundamental** science in the cloud
 - To ensure **repeatability** of research

The CloudLab Vision

- A “meta-cloud” for building clouds
- Build your own cloud on our hardware resources
- Agnostic to specific cloud software
 - Run existing cloud software stacks (like OpenStack, Hadoop, etc.)
 - ... or new ones built from the ground up
- Control and visibility all the way to the bare metal
- “Sliceable” for multiple, isolated experiments at

With CloudLab, it will be as easy to get a cloud tomorrow as it is to get a VM today

What Is CloudLab?

Slice A

*Geo-Distributed Storage
Research*

Slice B

*Stock
OpenStack*

- Supports transformative cloud research
- Built on Emulab and GENI
- Control to the bare metal
- Diverse, distributed resources
- Repeatable and scientific

Slice C

*Virtualization and
Isolation Research*

Slice D

*Allocation and Scheduling Research for
Cyber-Physical Systems*

Utah

Wisconsin

Clemson

GENI

CC-NIE, Internet2 AL2S, Regionals

CloudLab's Hardware

One facility, one account, three locations

- About 5,000 cores each (15,000 total)
- 8-16 cores per node
- Baseline: 4GB RAM / core
- Latest virtualization hardware
- TOR / Core switching design
- 10 Gb to nodes, SDN
- 100 Gb to Internet2 AL2S
- *Partnerships with multiple vendors*

Wisconsin

- **Storage and net.**
- Per node:
 - 128 GB RAM
 - 2x1TB Disk
 - 400 GB SSD
- Clos topology
- *Cisco*

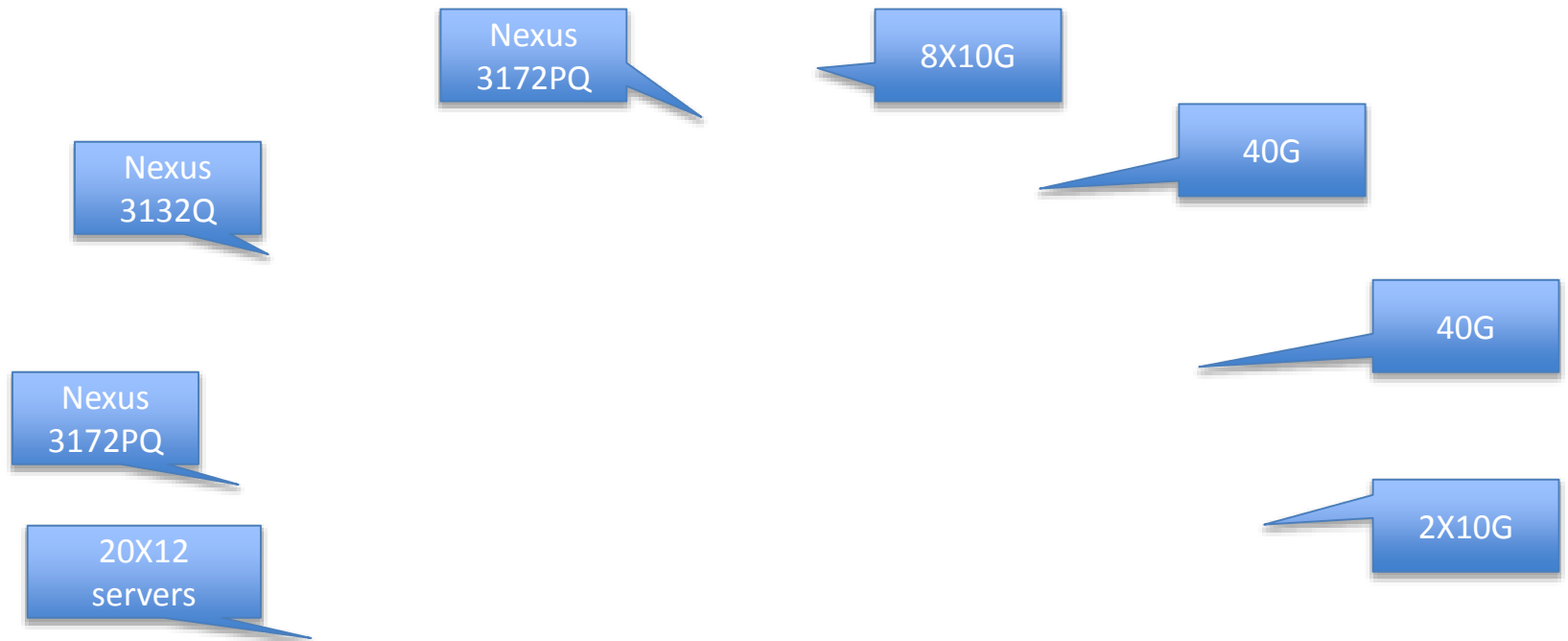
Clemson

- **High-memory**
- 16 GB RAM / core
- 16 cores / node
- Bulk block store
- Net. up to 40Gb
- High capacity
- *Dell*

Utah

- **Power-efficient**
- ARM64 / x86
- Power monitors
- Flash on ARM6s
- Disk on x86
- Very dense
- *HP*

Wisconsin/Cisco



Compute and storage

90X Cisco 220 M4



10X Cisco 240 M4



- 2X 8 cores @ 2.4GHz
 - 128GB RAM
 - 1X 480GB SSD
 - 2X 1.2 TB HDD
- 1X 1TB HDD
 - 12X 3TB HDD
(donated by Seagate)

Over the next year: ≥ 140 additional servers;
Limited number of accelerators, e.g., FPGAs, GPUs (planned)

Networking

Nexus 3132q



Nexus 3172pq



- OF 1.0 (working with Cisco on OF 1.3 support)
- Monitoring of instantaneous queue lengths
- Fine-grained tracing of control plane actions
- Support for multiple virtual router instances per router
- Support for many routing protocols

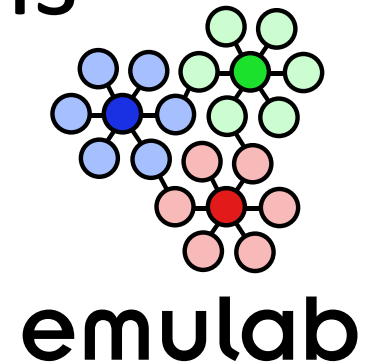
Experiments supported

Large number of nodes/cores, and bare-metal control over nodes/switches, for sophisticated network/memory/storage research

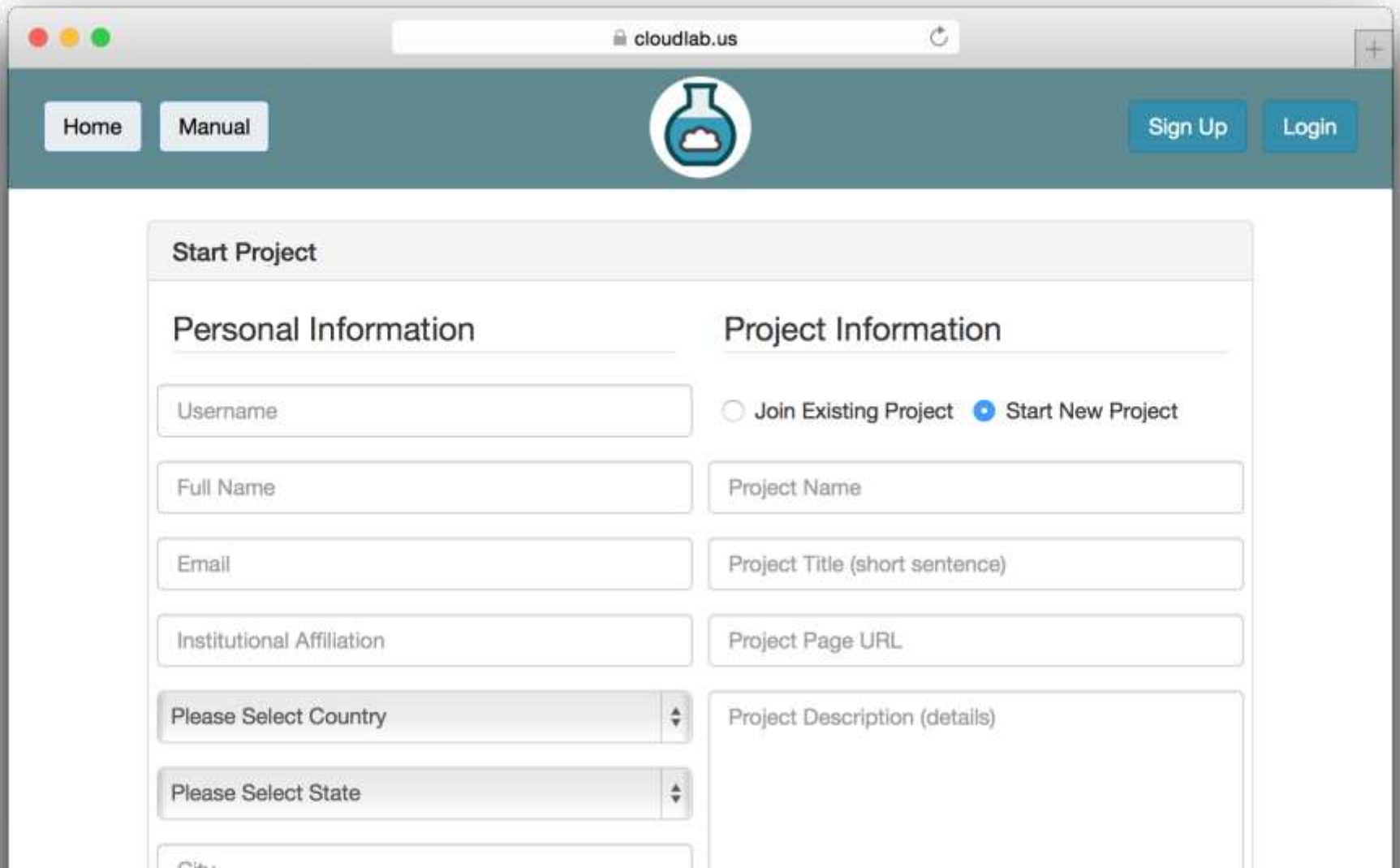
- ... Network I/O performance (e.g., Presto), intra-cloud routing (e.g., Conga) and transport (e.g., DCTCP)
- ... Network virtualization (e.g., CloudNaaS, OpenNF)
- ... In-memory big data frameworks (e.g., Spark/Shark, Graphene)
- ... Cloud-scale resource management and scheduling (e.g., Mesos; Tetris)
- ... New models for Cloud storage (e.g., tiered; flat storage; IOFlow, split-level scheduling)
- ... New architectures (e.g., RAMCloud for storage)

Technology Foundations

- Built on Emulab and GENI (“ProtoGENI”)
- Provisions, then gets out of the way
 - “Run-time” services are optional
- Controllable through a web interface and GENI APIs
- *Scientific instrument for repeatable research*
 - Physical isolation for most resources
 - *Profiles* capture everything needed for experiments
 - Software, data, and hardware details
 - Can be shared and published (eg. in papers)



Sign Up At CloudLab.us



A screenshot of a web browser window showing the CloudLab.us sign-up page. The browser's address bar displays 'cloudlab.us'. The page has a dark teal header with 'Home' and 'Manual' links on the left, a logo in the center, and 'Sign Up' and 'Login' buttons on the right. The main content area is titled 'Start Project' and is divided into two columns: 'Personal Information' and 'Project Information'. The 'Personal Information' column contains fields for Username, Full Name, Email, Institutional Affiliation, a Country dropdown menu, a State dropdown menu, and a City field. The 'Project Information' column contains radio buttons for 'Join Existing Project' and 'Start New Project' (which is selected), a Project Name field, a Project Title (short sentence) field, a Project Page URL field, and a larger Project Description (details) text area.

cloudlab.us

Home Manual

Sign Up Login

Start Project

Personal Information

Username

Full Name

Email

Institutional Affiliation

Please Select Country

Please Select State

City

Project Information

☐ Join Existing Project ☒ Start New Project

Project Name

Project Title (short sentence)

Project Page URL

Project Description (details)

