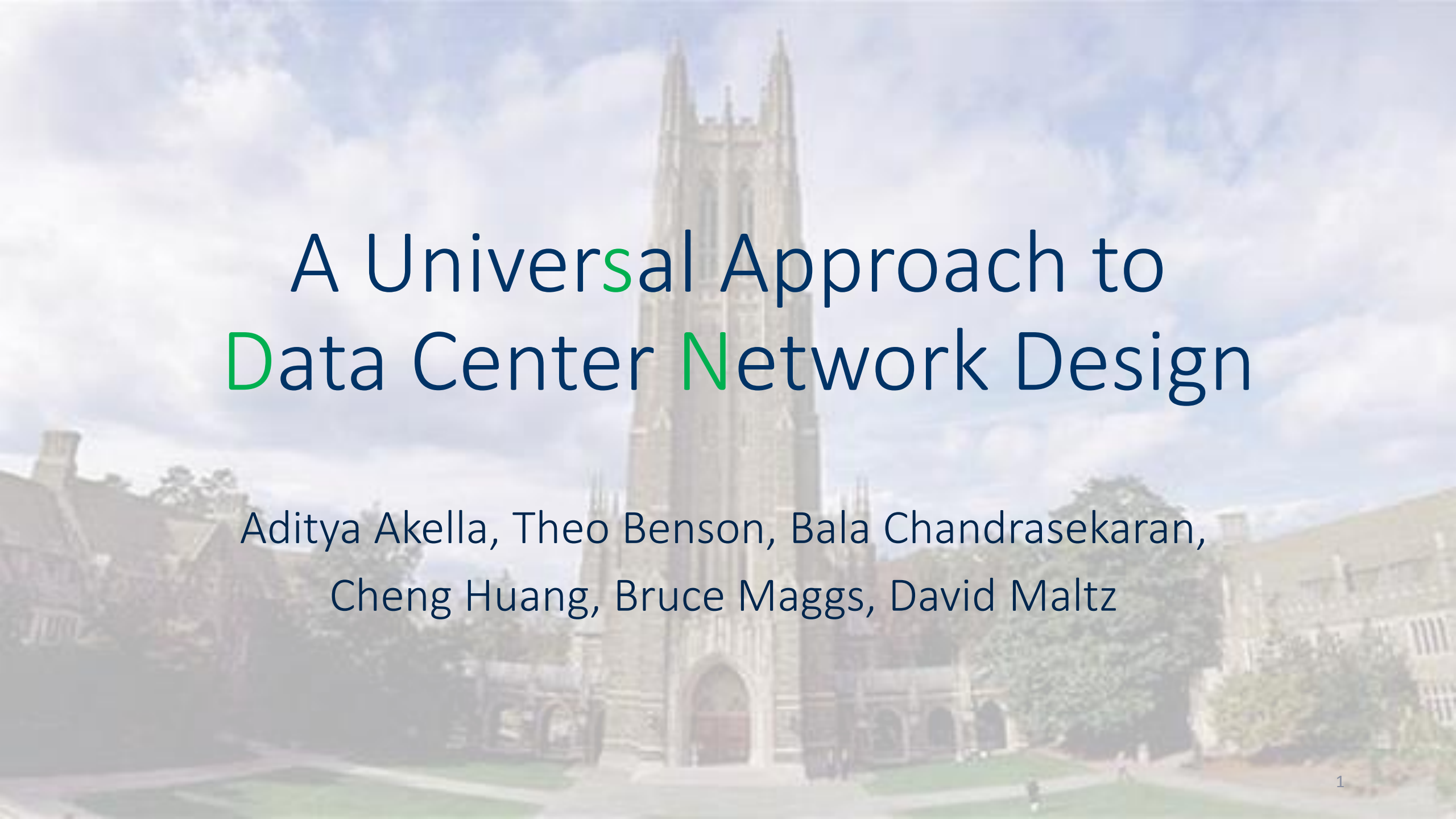
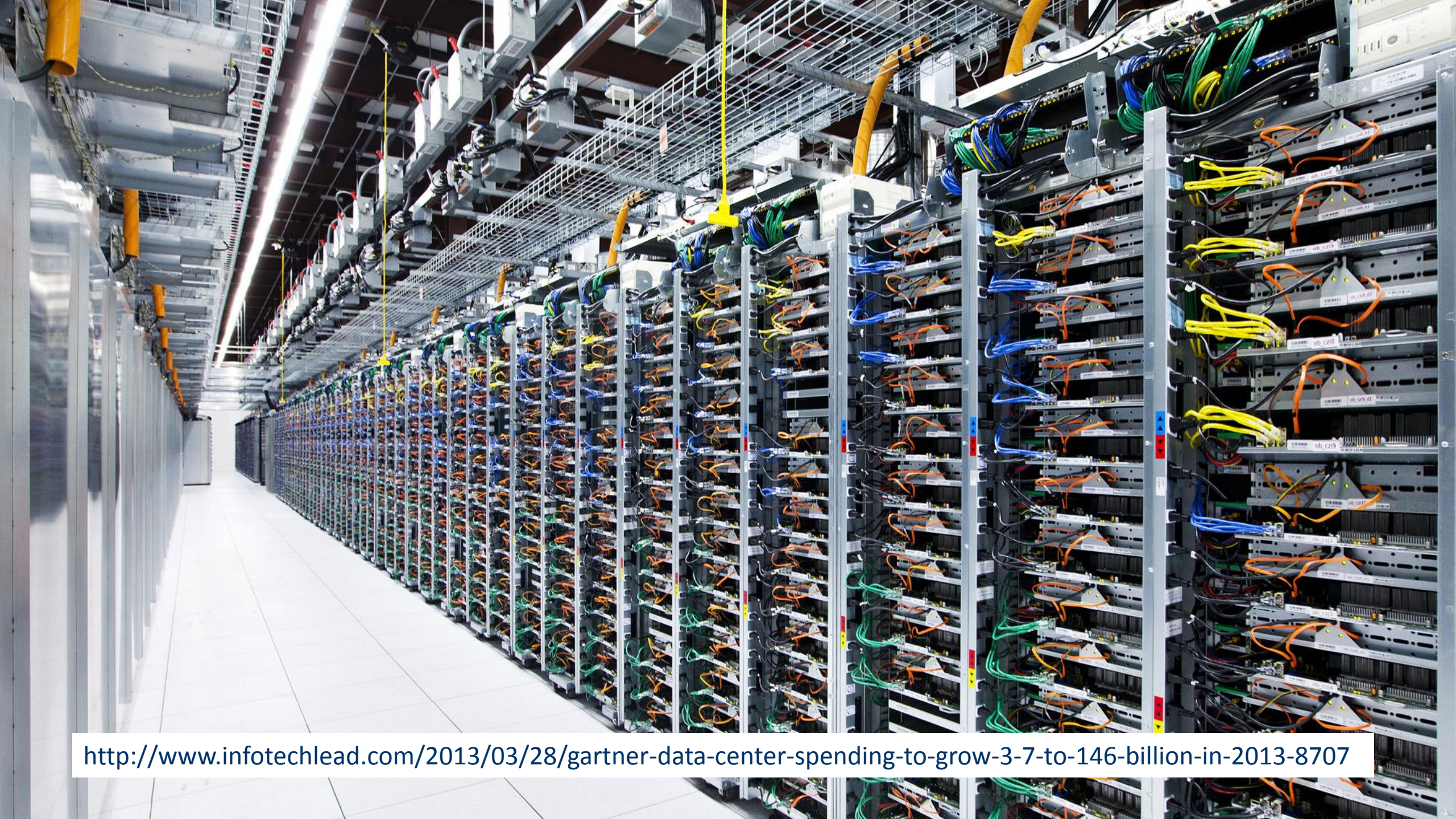


A Universal Approach to Data Center Network Design

Aditya Akella, Theo Benson, Bala Chandrasekaran,
Cheng Huang, Bruce Maggs, David Maltz



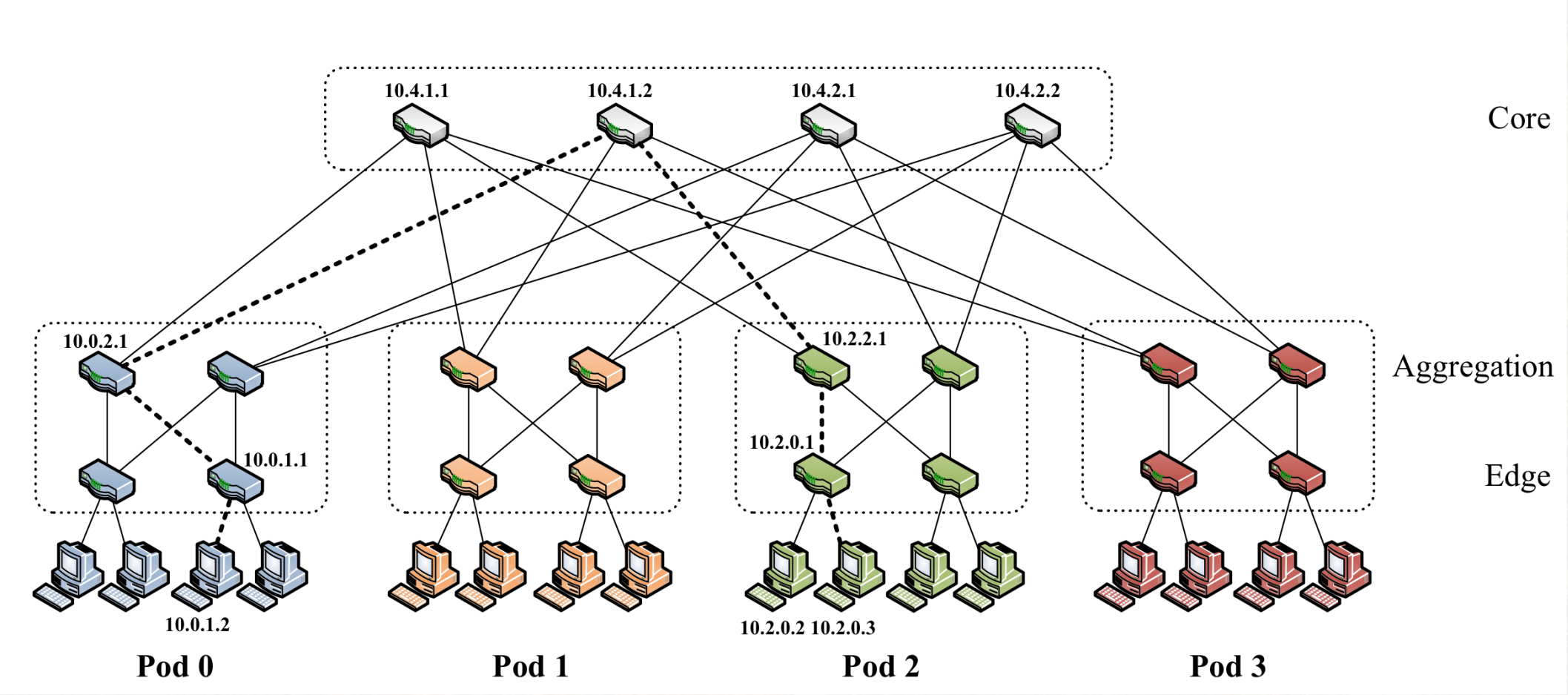
http://www.infotechlead.com/2013/03/28/gartner-data-center-spending-to-grow-3-7-to-146-billion-in-2013-8707

What to build?

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

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

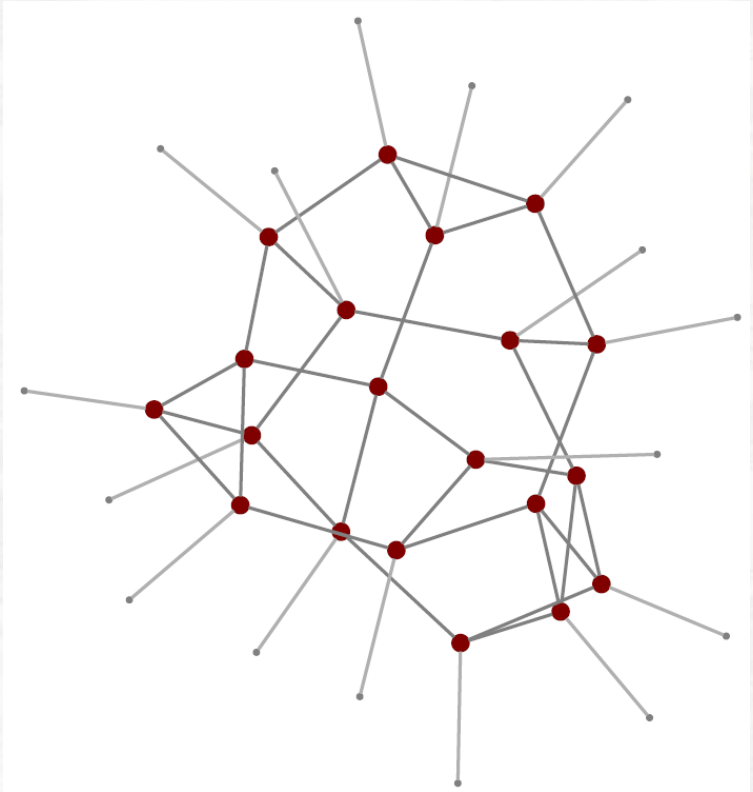
“Fat-tree” SIGCOMM 2008



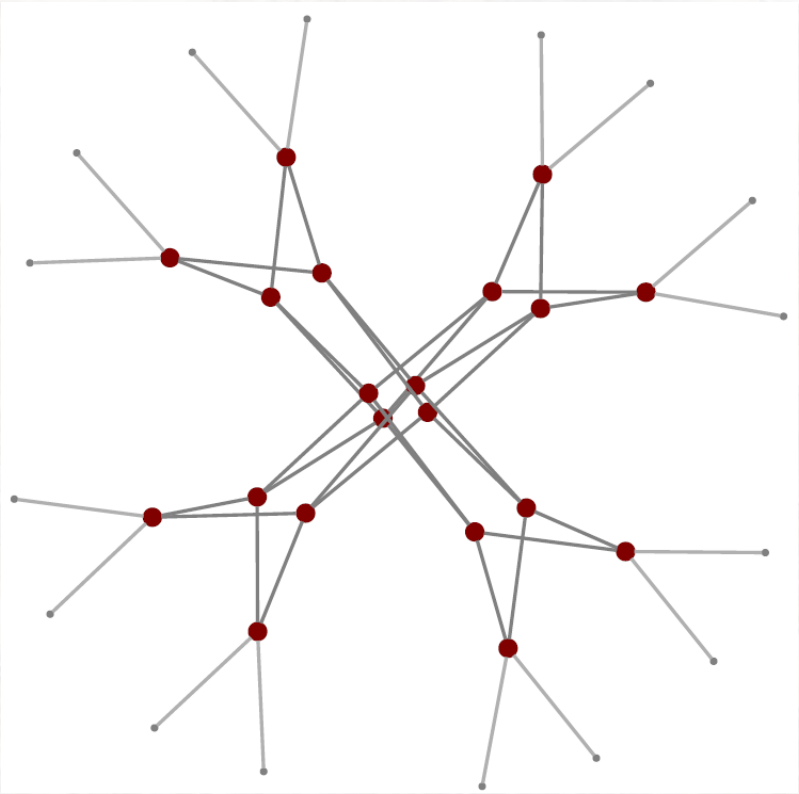
Isomorphic to butterfly network except at top level

Bisection width n/2, oversubscription ratio 1

Jellyfish (NSDI 2012)

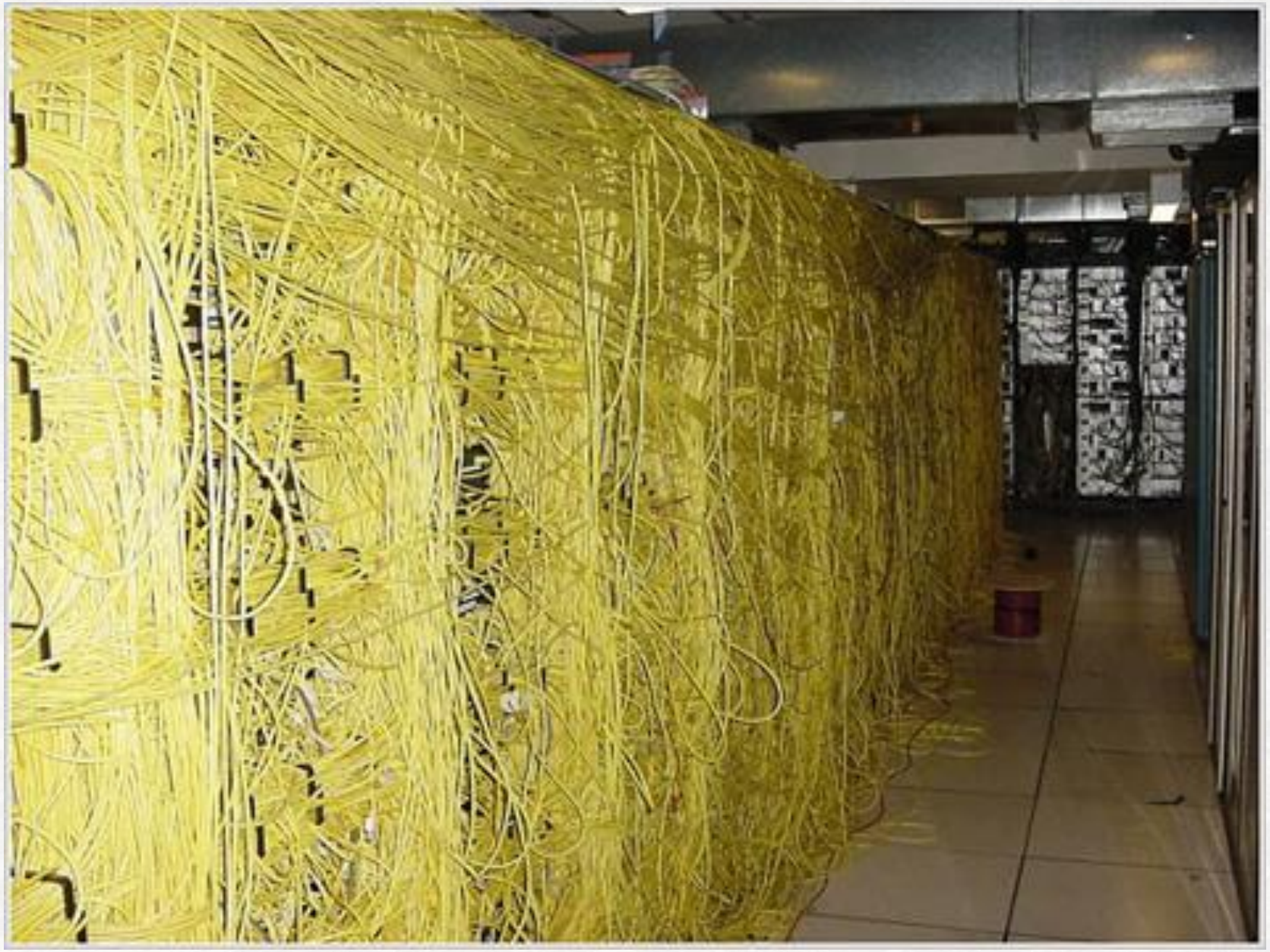


vs.



Random connections
Bisection width Θ(n)

~~"fat-tree"~~ butterfly



Duke UNIVERSITY

http://www.itdisasters.com/category/networking-rack/

How to compare networks?

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

A Universal Approach

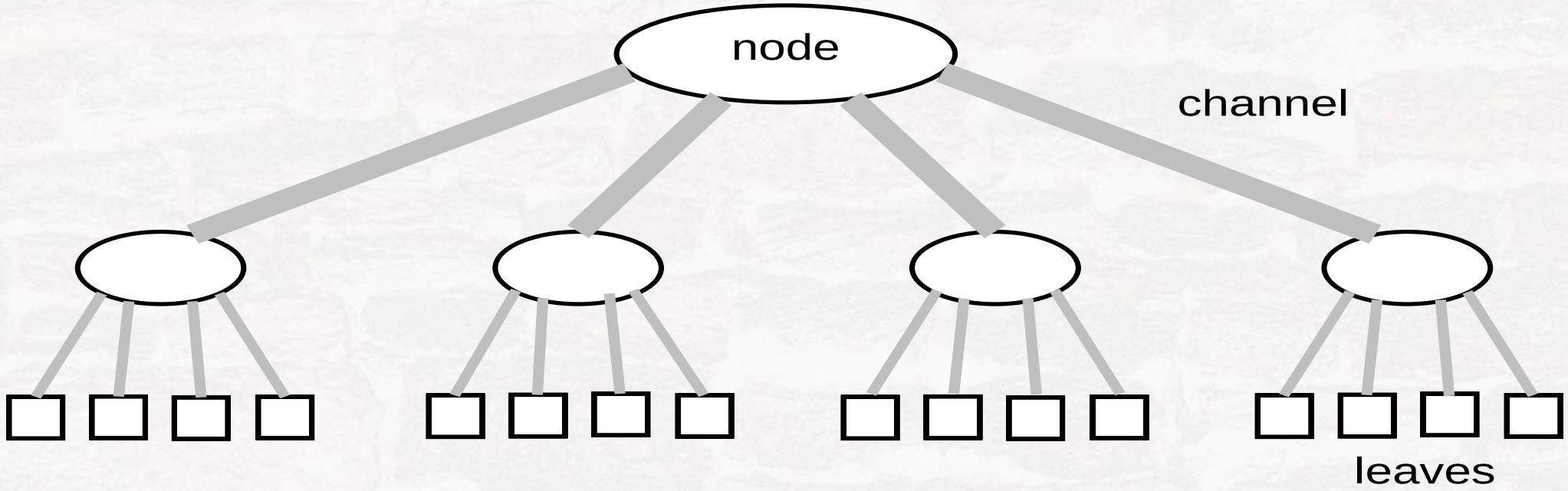
Build a single network that is competitive, for any application, with any other network that can be built at the same cost.

Area-Universality

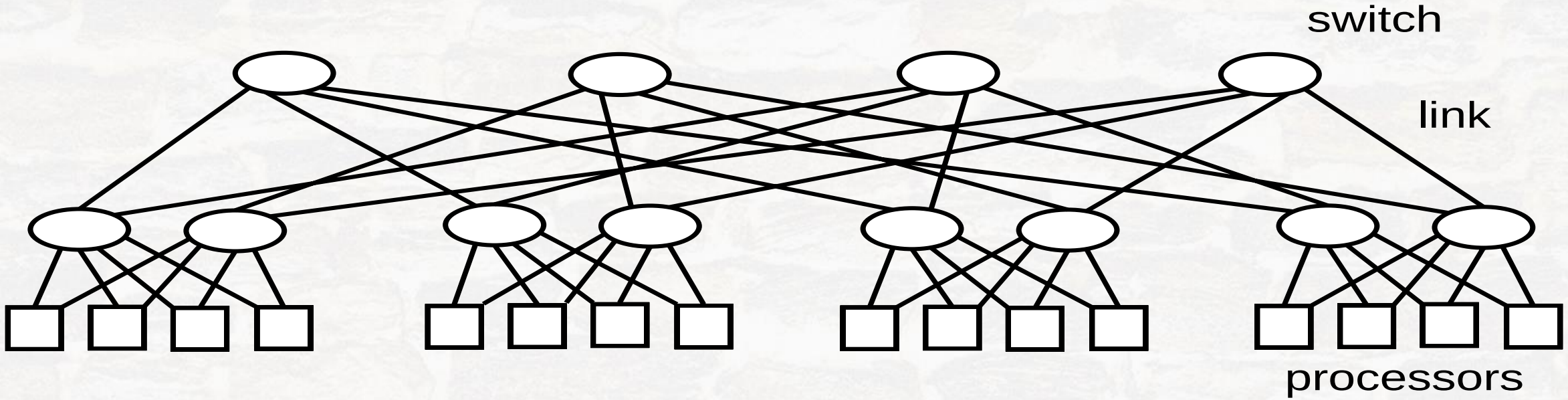
Theorem [Leiserson, 1985]: There is a fat-tree network of area n that can emulate any other network that can be laid out in area n with slowdown $O(\log^3 n)$.

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

Coarse Structure of Fat-Tree Network

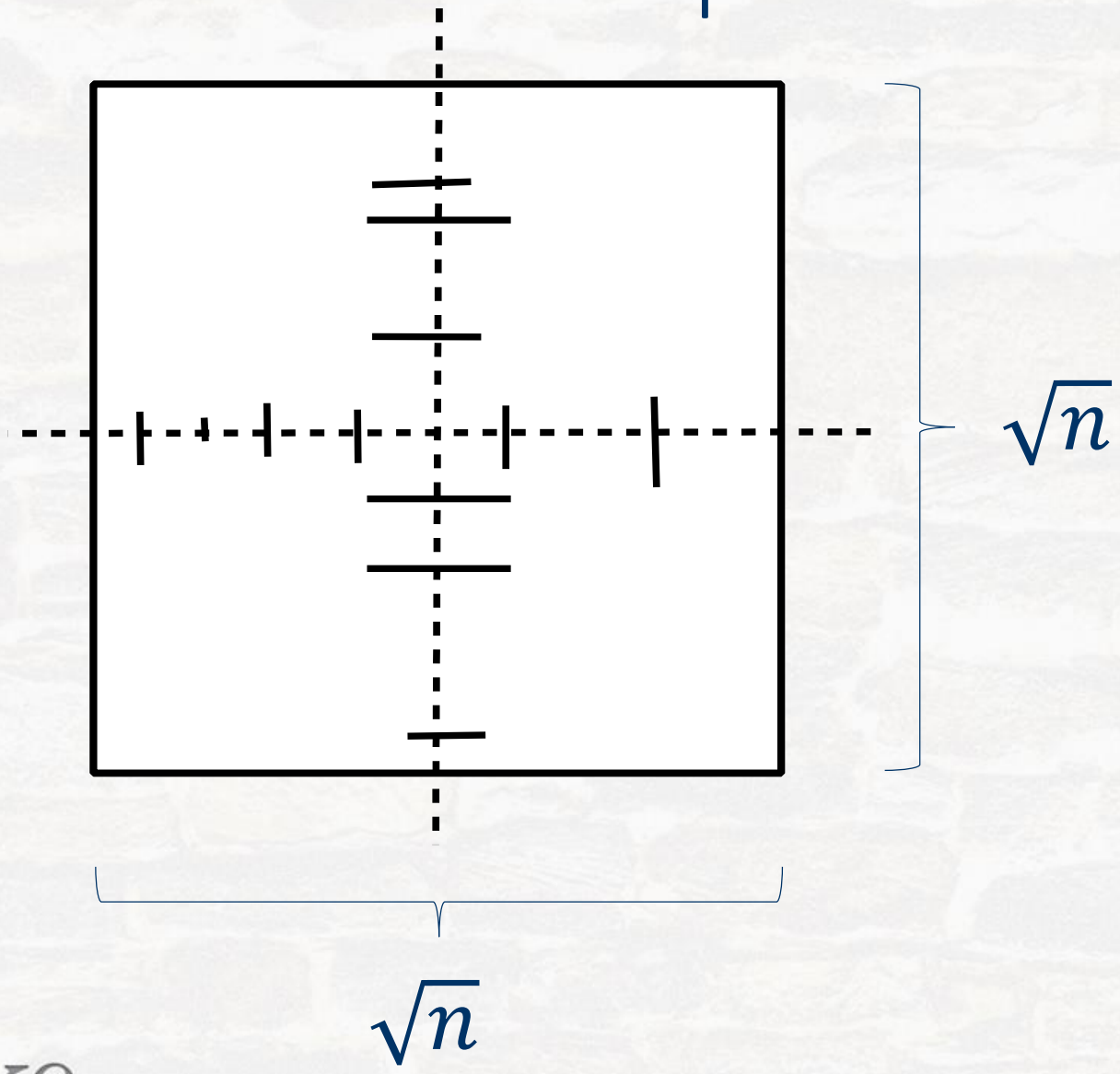


Example of Fine Structure of a Fat-Tree



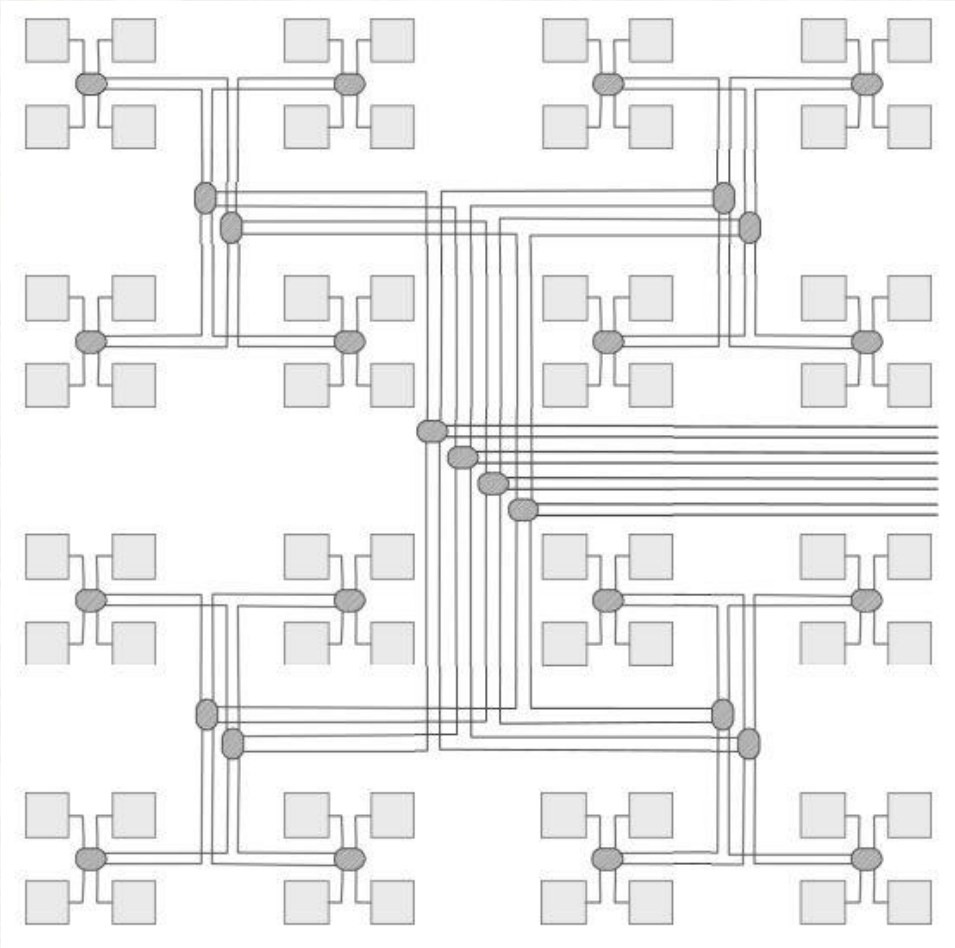
Butterfly Fat-Tree (Greenberg-Leiserson)

Recursive Decomposition of VLSI Layout

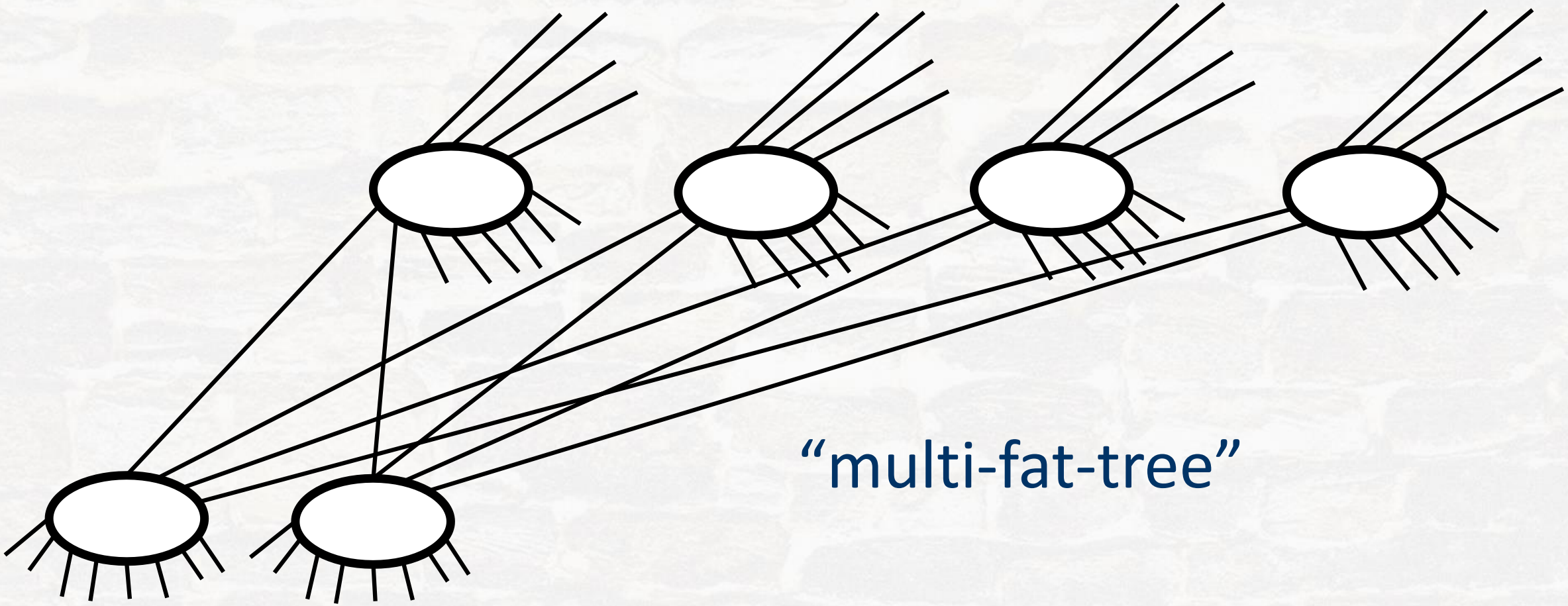


Idea: match fat-tree channel capacity with maximum number of wires cut in layout.

Layout of Area-Universal Fat-Tree



Channel with Redundant Links



"multi-fat-tree"

No Magic Formula

[Leiserson 1989] In practice, of course, no mathematical rule governs interconnect technology. Most networks that have been proposed for parallel processing … are inflexible when it comes to adapting their topologies to the arbitrary bandwidths provided by packaging technology… The channels of a fat-tree can be adapted to effectively utilize whatever bandwidths the technology can provide and which make engineering sense in terms of cost and performance.

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.

No special-purpose network can allocate more than a factor of three or four more at any level.

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 (Ethernet cable has diameter 0.54cm)
- Assumes that bandwidth, not latency, limits performance.

Costs (2014)

- A switch costs approximately $125 per 10Gbps port
- 10Gbps Direct Attached Copper (DAC) line (5m) is $50-$100
- 10Gbps fiber with optical modules (20m) is around $200

Trends

Looked at large-scale data center 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.