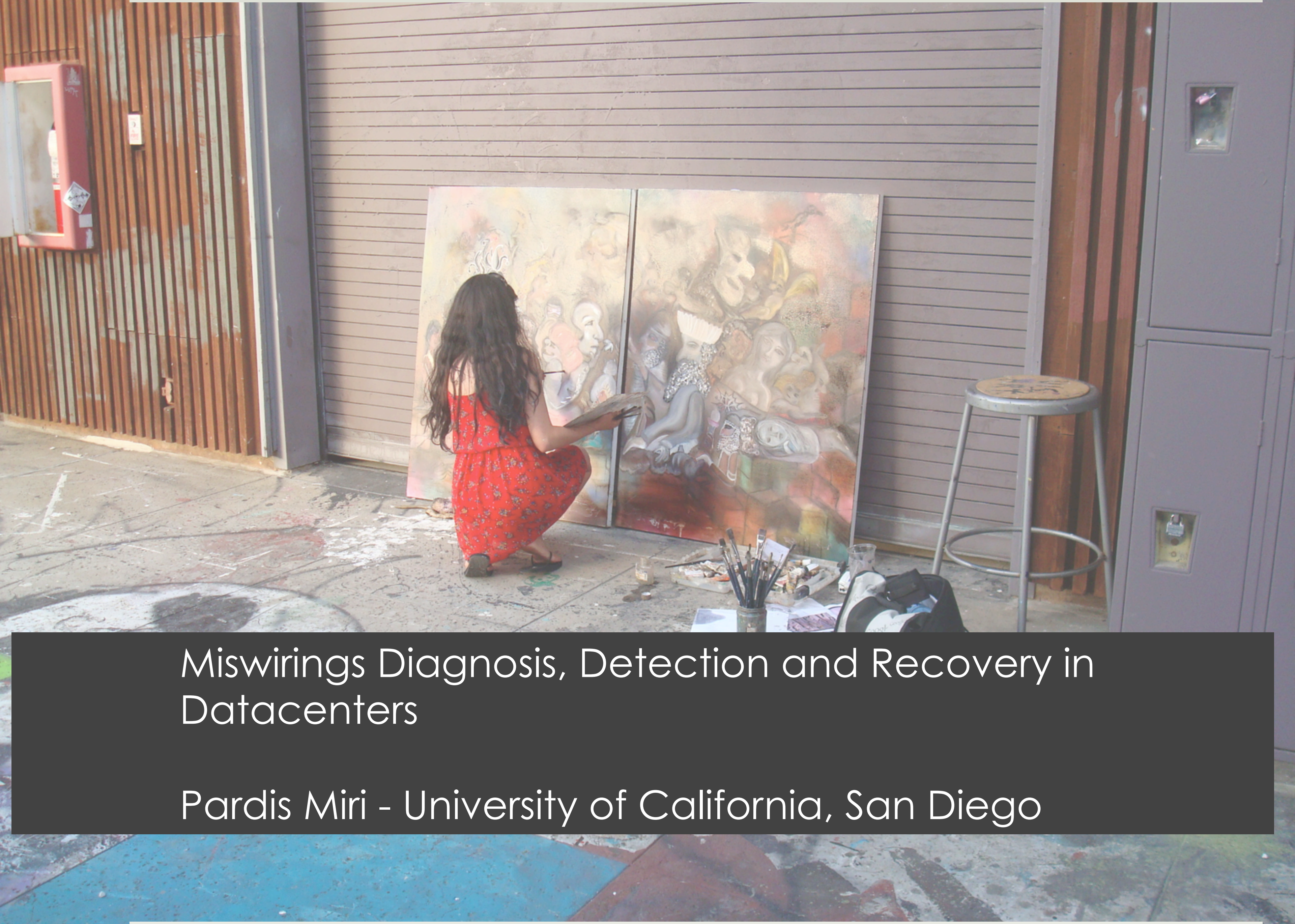


Miswirings Diagnosis, Detection and Recovery in Datacenters

Pardis Miri - University of California, San Diego

Who is the Audience of this talk?

Executive Overview

- What is big data?
  - We create 2.5 petabytes of data every day.
    - Sensors used to gather climate info.
    - Social media sites
    - digital pictures and videos.
  - Big data is a BIG concern!

Executive Overview

- Where is Big data sitting?
  - On Datacenters

Executive Overview

- What is a Datacenter?
  - Switches
  - Server Clusters
  - Cables
  - Racks
  - Power Supply
  - Air Conditioning
  - Big Space

Executive Overview

- **Cabling** is a big issue! It might kill performance.
  - Incorrect cabling
  - Machine failures
  - Component and partitions add & removal
- **Scalability** of a datacenter architecture is another BIG issue!

What is Datacenter?

- Microsoft Chicago Datacenter
- 700,000 sf – 16 football fields
- 60 megawatts power
- As big data grows,

  datacenter must scale.

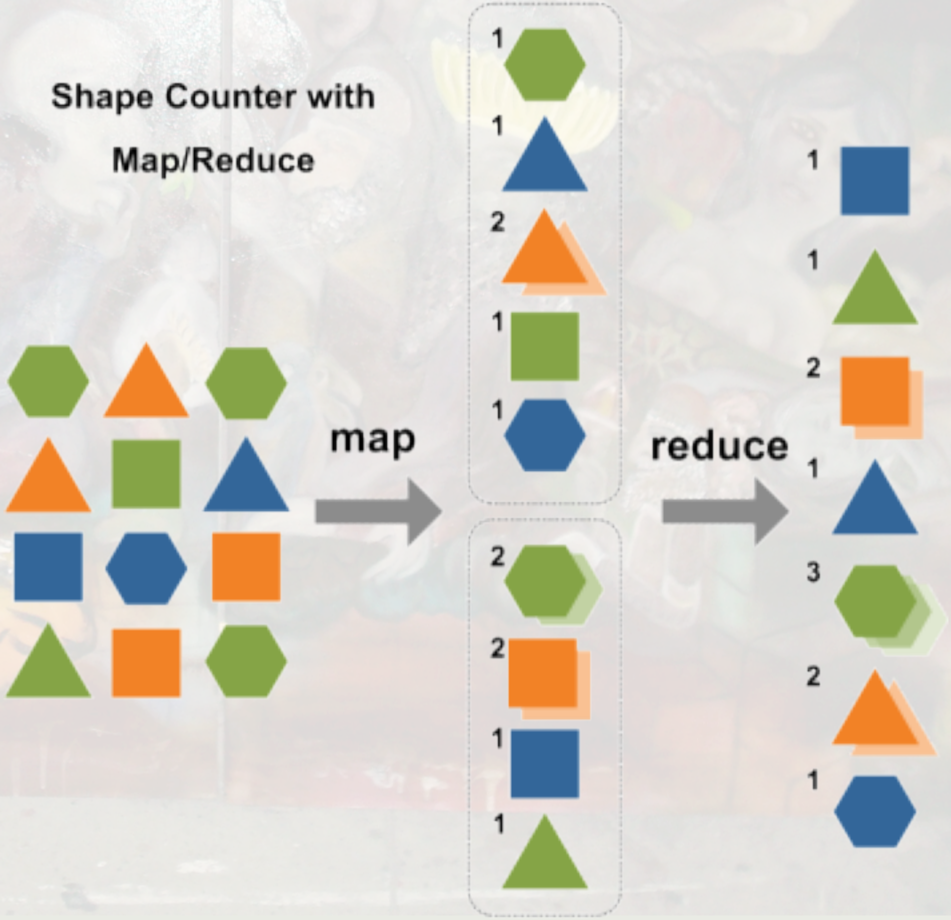
Microsoft's Chicago Data Center
October 2008
Photo By McShane Fleming Studios, Chicago

Executive Overview

- What is big data?
  - Data sets ranging from a few dozen terabytes to many petabytes. Examples: web logs, social networks and etc.
- Where Big data is sitting? On a Datacenter
- What is a datacenter?
  - Switches and end-host servers wired up together sitting in racks.
- **How Big Data is managed on a Datacenter?**
  - **By Running MapReduce**

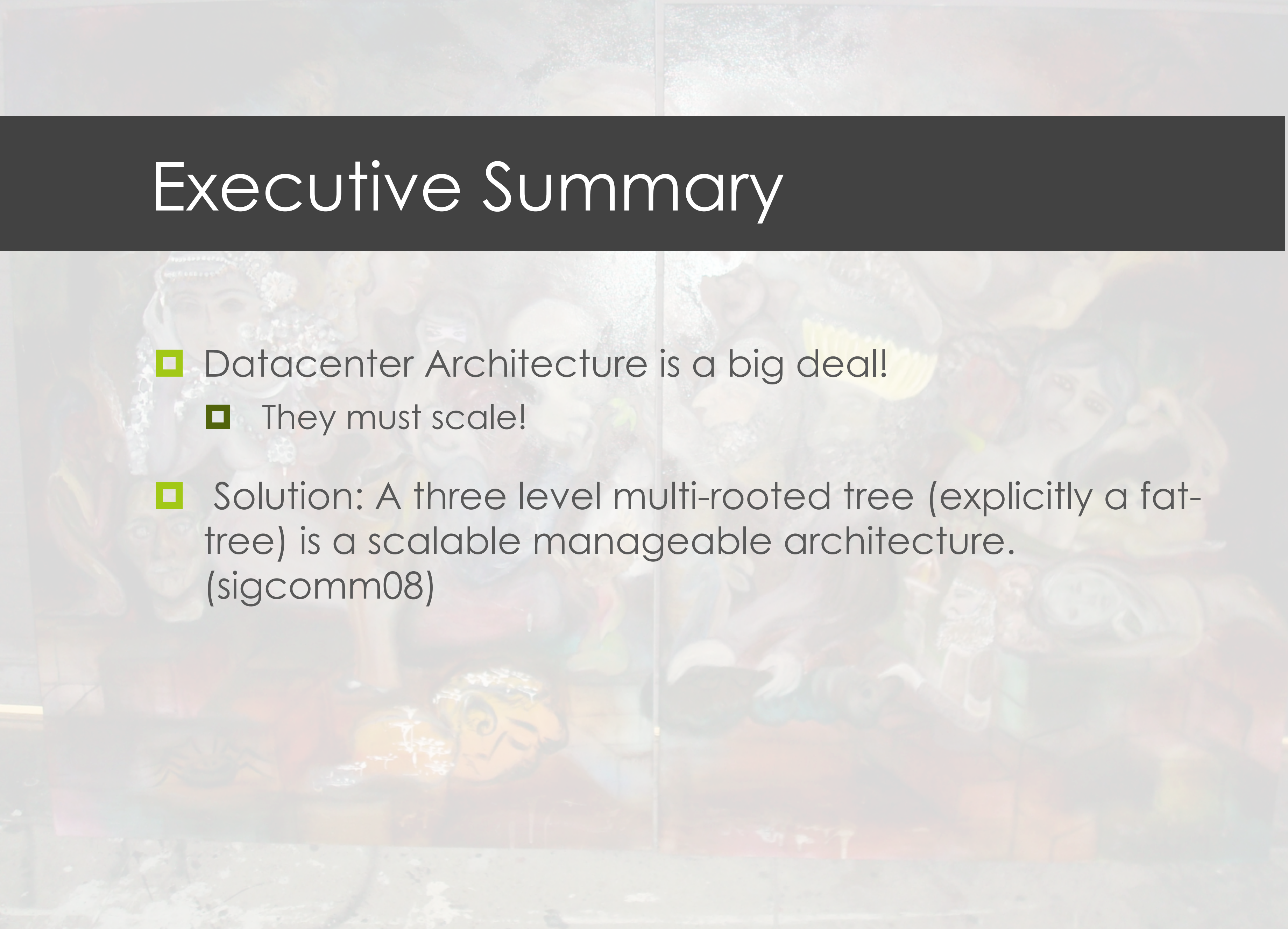
What is MapReduce?

- The caller Maps to two nodes.
- The nodes perform the counting and return the results to the caller
- The caller reduces the results and announces them.



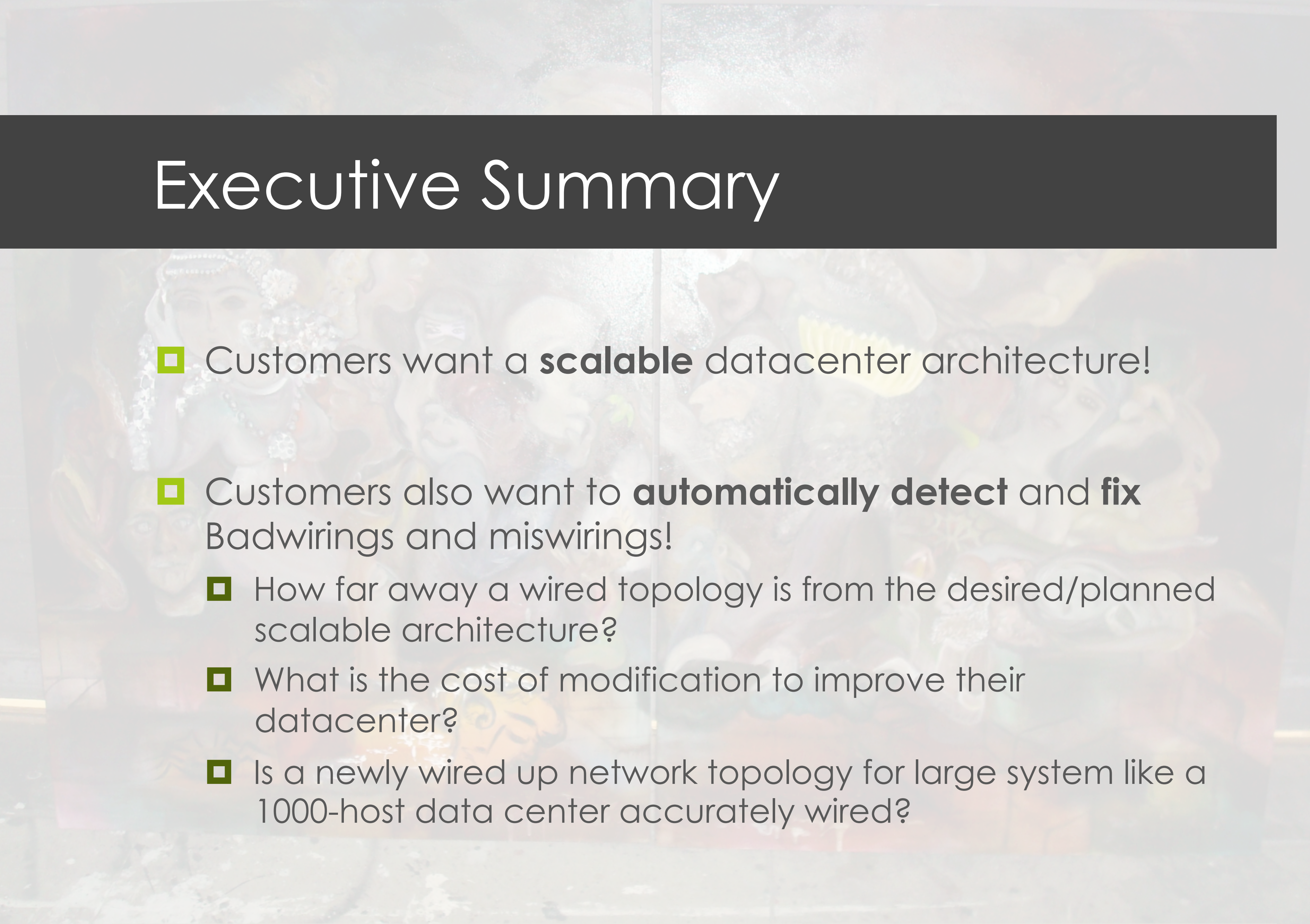
Executive Overview

- What is big data?
  - Data sets ranging from a few dozen terabytes to many petabytes. Examples: web logs, social networks and etc.
- Where Big data is sitting? On a Data Center
- What is a data center?
  - Switches and end-host servers wired up together sitting in racks.
- How Big Data is managed on a Data center?
  - By Running MapReduce
- What is MapReduce?
  - Programing Model to destitute work (receiving and retrieving data) across a datacenter.



Executive Summary

- Datacenter Architecture is a big deal!
  - They must scale!
- Solution: A three level multi-rooted tree (explicitly a fat-tree) is a scalable manageable architecture. (sigcomm08)



Executive Summary

- Customers want a **scalable** datacenter architecture!
- Customers also want to **automatically detect** and **fix** Badwirings and miswirings!
  - How far away a wired topology is from the desired/planned scalable architecture?
  - What is the cost of modification to improve their datacenter?
  - Is a newly wired up network topology for large system like a 1000-host data center accurately wired?

Executive Summary

- So, customers want a **diagnostic protocol** to detect badwirings and miswirings that returns a list of changes to improve performance.
  - Example of list of changes:
    - Remove host x from switch s.
    - Remove the cable connecting switch x and switch y through ports x' and y'.
    - add a cable between switches x and y.
    - replace switch x with a larger switch that supports at least z ports.
    - swap cables a and b connecting certain switches.
- **This talk is about development of such a diagnostic protocol.**

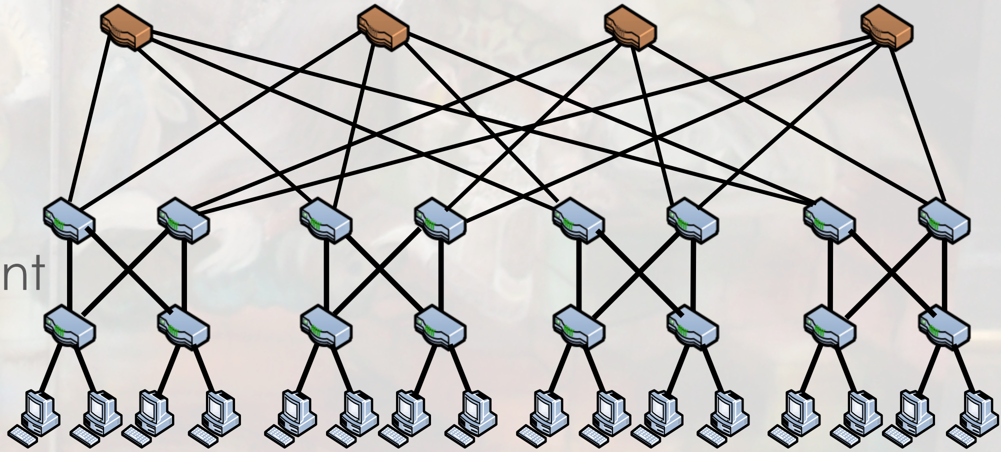
The Rest of the Talk Schedule

- Datacenter Architecture
  - fat-tree and its properties
    - What is over-subscription ratio?
    - What are they types of mis/bad-wirings?
  - What algorithms are we using to detect each miss/bad-wirings?

Fat-tree: Desirable Datacenter Architecture

- So, the magical desired **scalable** datacenter Architecture is a 3-level multi-rooted tree (fat-tree).

- It is ….
  - scalable
  - cost efficient
  - multi-path routable
  - reasonably fault tolerant



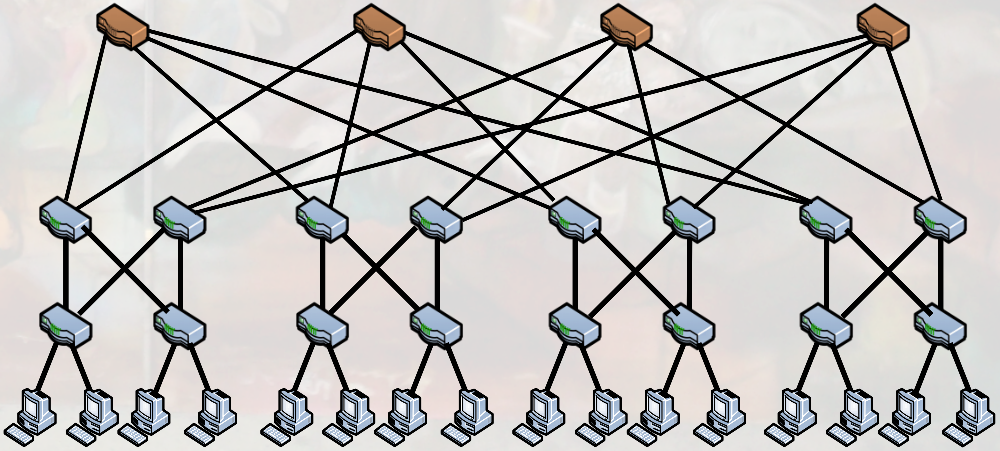
- **PortLand: A Scalable Fault-Tolerance Layer 2 Data Center Network Fabric.** *Niranjan Mysore et al.* SIGCOMM 2009

Fat-tree: Desirable Datacenter Architecture

- Embedding routing info in the MAC address made this architecture scalable.
- Switches have limited forwarding entries.
- Mac address is a unique in use in L2.
  - Datacenter architecture cannot scale in L2.
- Manual configuration of switches does not scale in L3.
- Using pseudo-MAC address made this architecture scalable.

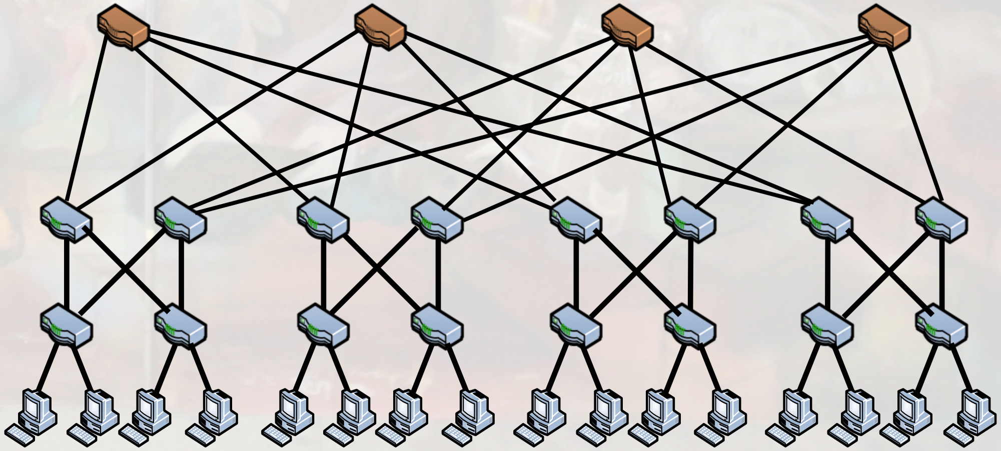
Fat-tree: Desirable Datacenter Architecture

- 16 end-host fat-tree built with 4-port switches:
  - Scalable
    - K-port switches
    - $\frac{K^3}{4}$ end-hosts
    - $\frac{5K^2}{4}$ switches
    - K switches in a putative pod



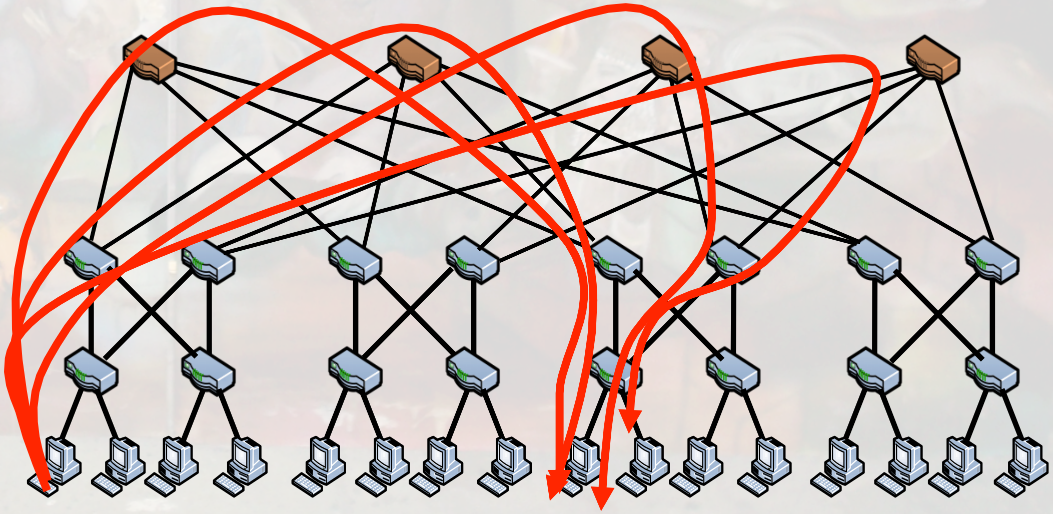
Fat-tree: Desirable Datacenter Architecture

- 16 end-host fat-tree built with 4-port switches:
  - Cost efficient
    - Reasonable switch size (K= 48)
    - End-hosts = 27K



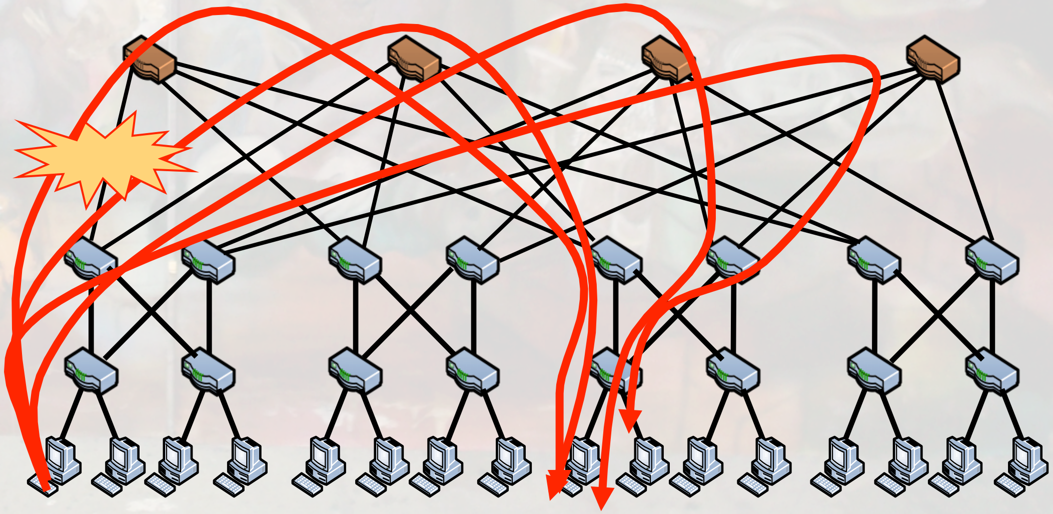
Fat-tree: Desirable Datacenter Architecture

- 16 end-host fat-tree built with 4-port switches:
  - multi-path routes
    - $\frac{K^3}{6}$ Routes



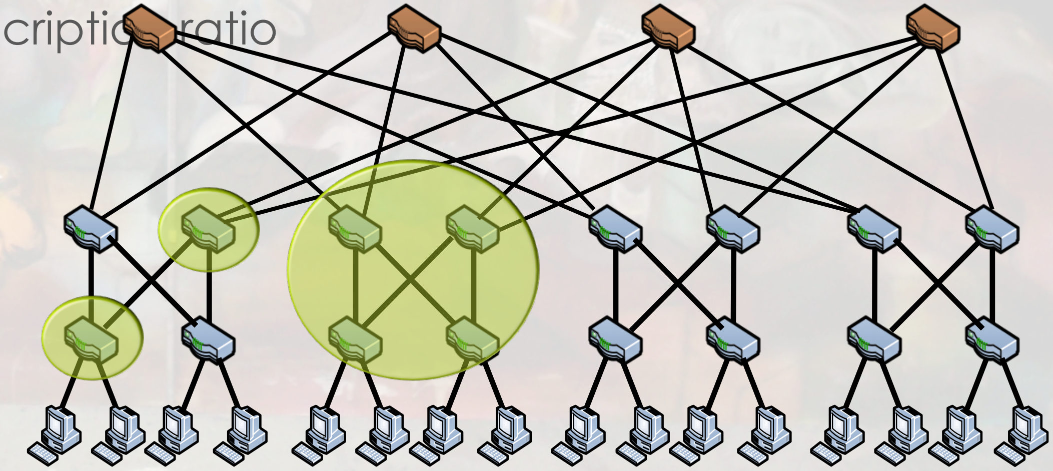
Fat-tree: Desirable Datacenter Architecture

- 16 end-host fat-tree built with 4-port switches:
  - Fault tolerance



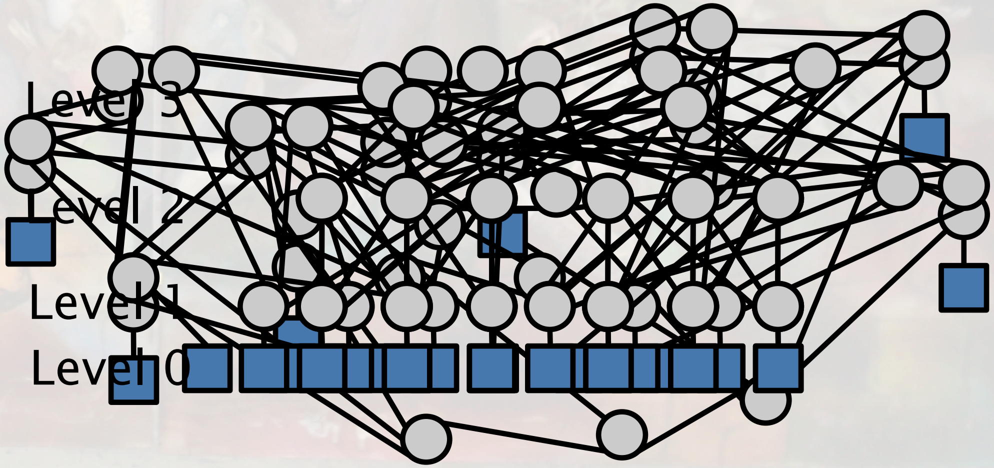
Fat-tree: Desirable Datacenter Architecture

- 16 end-host fat-tree built with 4-port switches:
  - with reasonable **oversubscription ratio**
  - **OR** = $\frac{accessBandwidth}{guaranteedBandwith}$
  - intra-pod oversubscription ratio



Level Assignment Protocol

- Protocol that detects the 3 levels of a multi-rooted tree.



Core Level 3

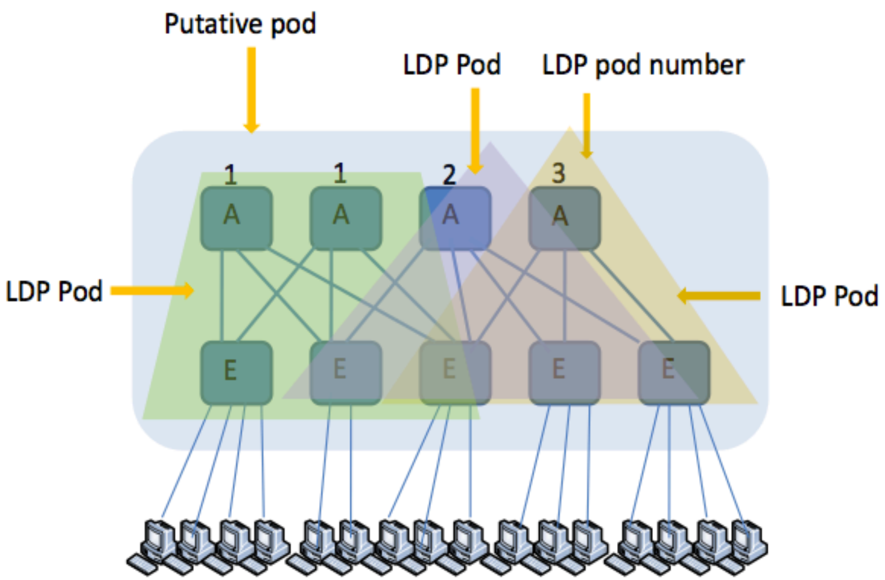
Aggregation Level 2

Edge Level 1

End Hosts Level 0

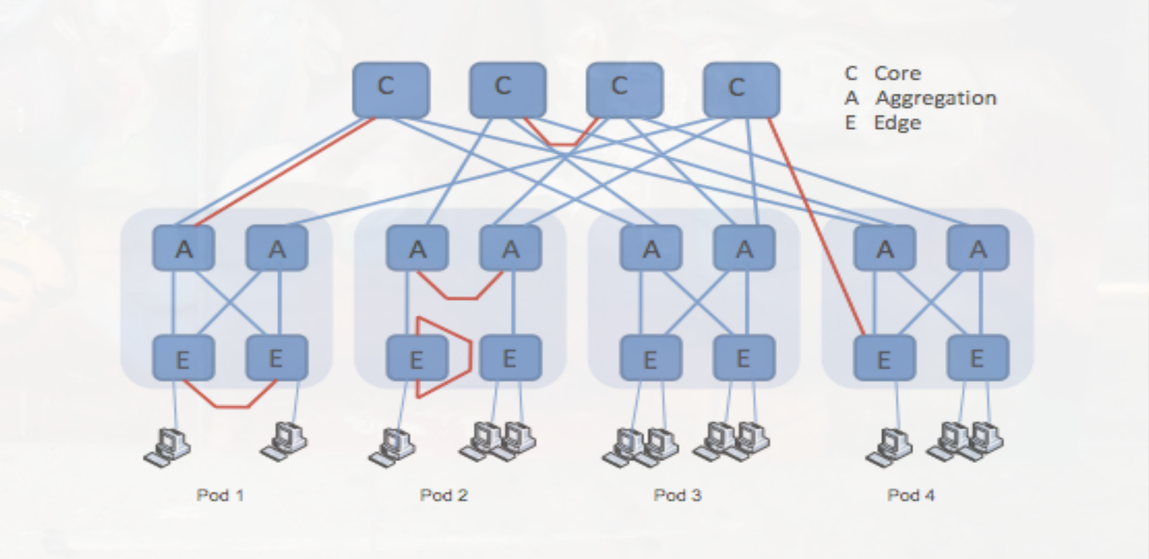
LDP and Putative Pods

- Putative Pod: all the switches that can be reached via a BFS or DFS algorithm.
- LDP Pod: all edge switches sharing the same set of aggregation switches + those aggregation switches



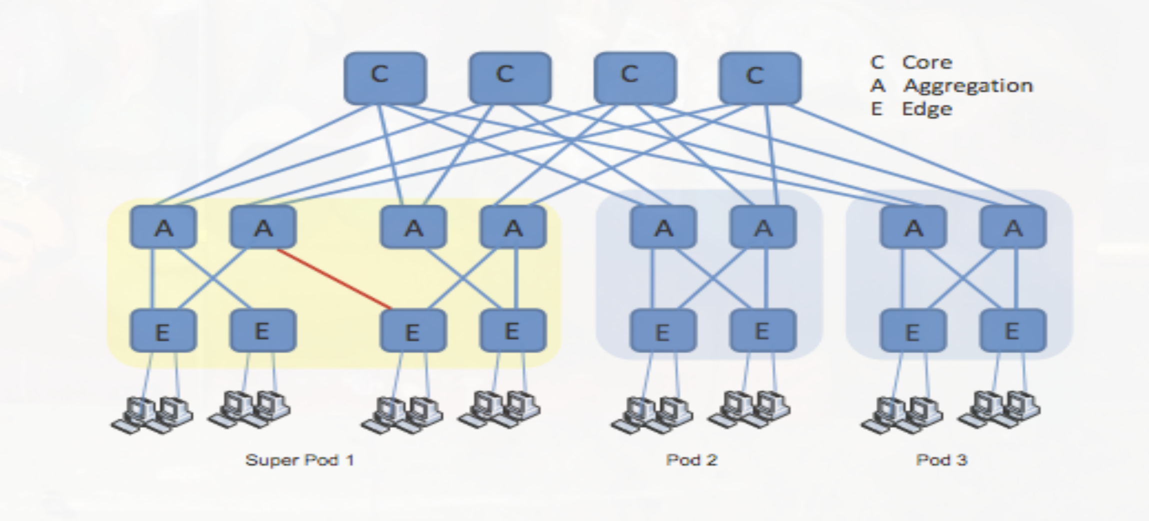
Diagnostic Protocol Phases

- Phase 1: Detection and deactivation of the following miswirings
  - E-C
  - E-E
  - A-A
  - C-C
  - A-C
  - Loopback



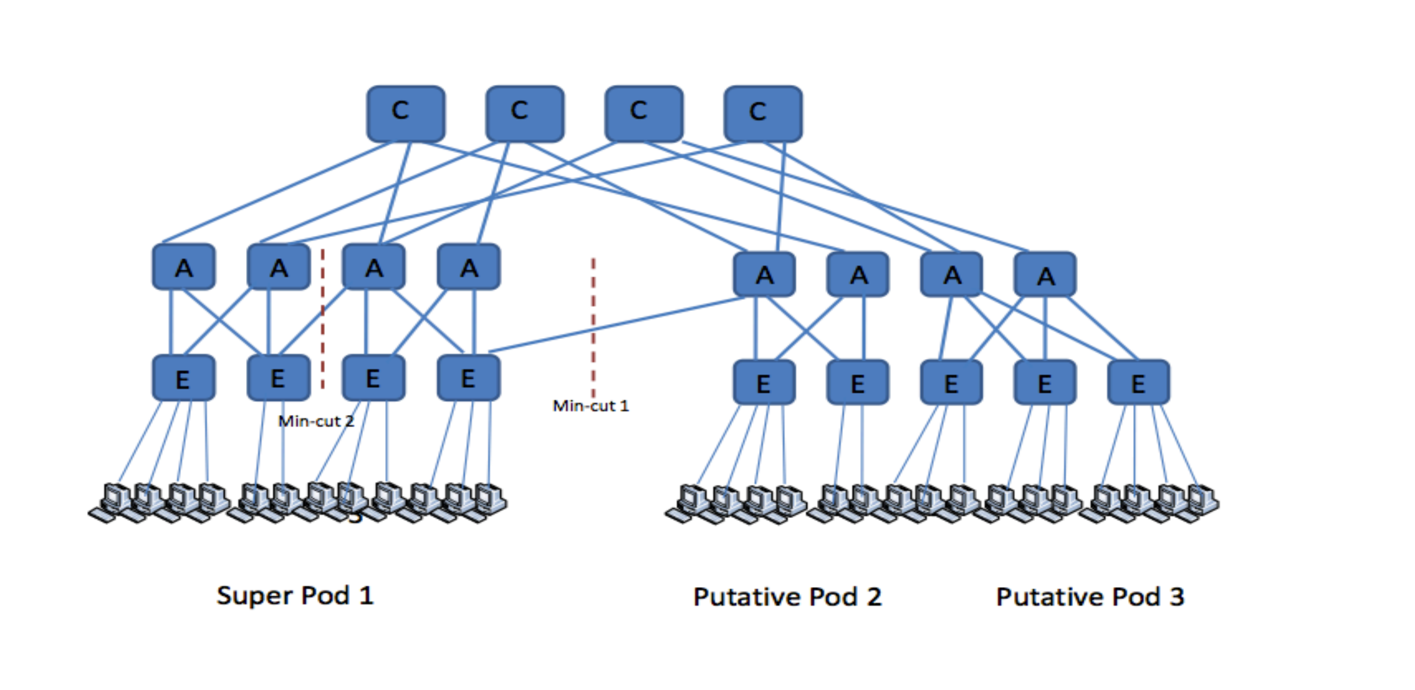
Types of Fat-tree Miswirings

- Phase 2: Detect and deactivate links that cause Super pod formation
  - Number of switches within a pod is greater than K then super pod is formed.
  - Solution: Min-cut algorithm to find the Cut(s)!



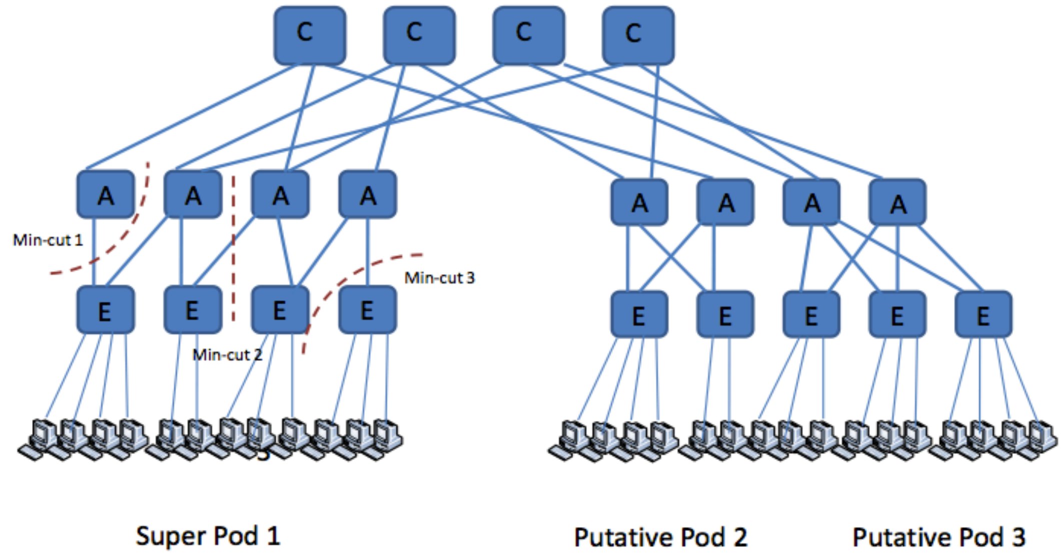
Types of Fat-tree Miswirings

- Our Min-cut algorithm discovers all the cuts.
  - Check the validity of the cut.



Types of Fat-tree Miswirings

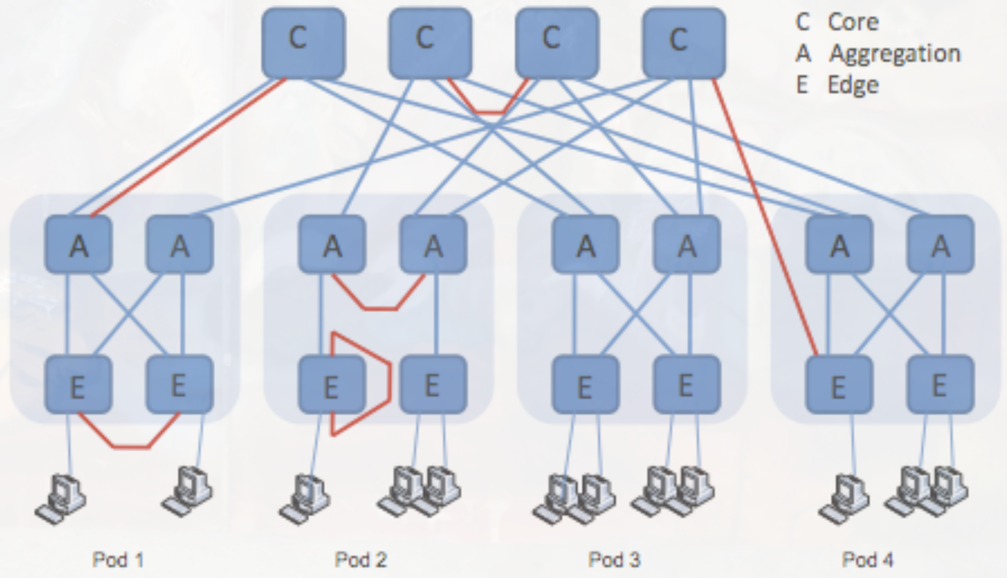
- Min-cut 2 is valid.
- Min-cut 1 and 3 are invalid.



Types of Fat-tree Miswirings

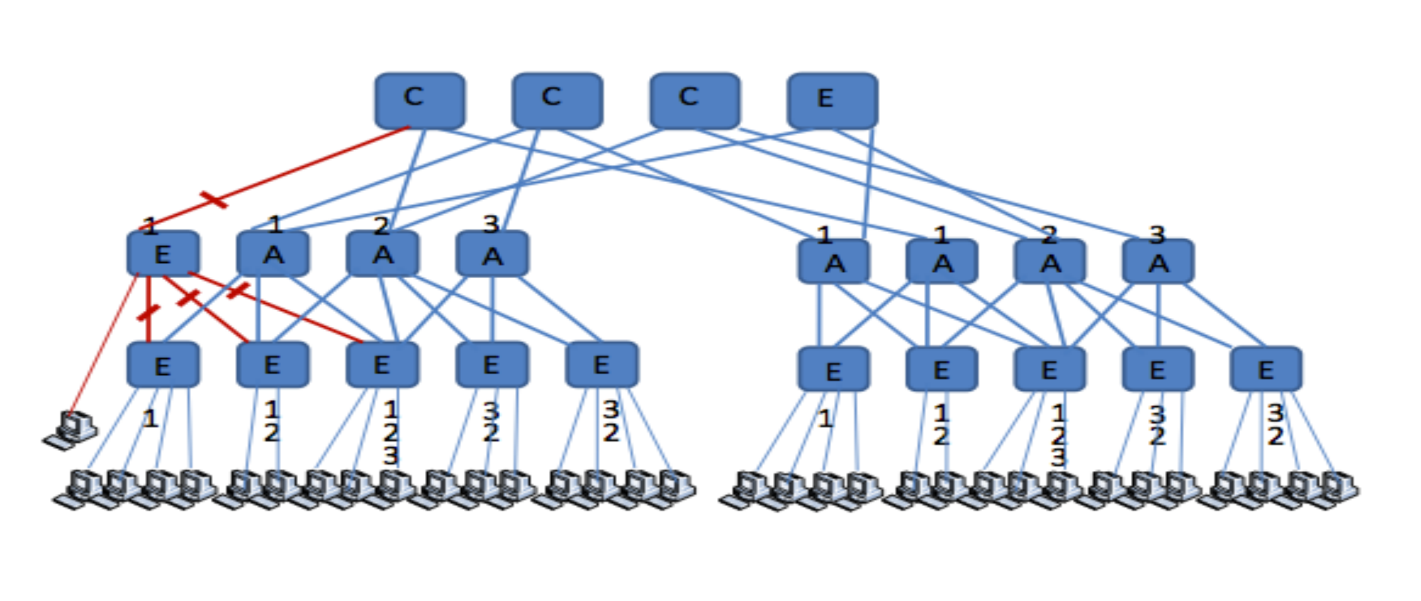
- Group 1: Six cases of badwiring

- E-E
- A-A
- C-C
- A-C
- E-C
- Loopback



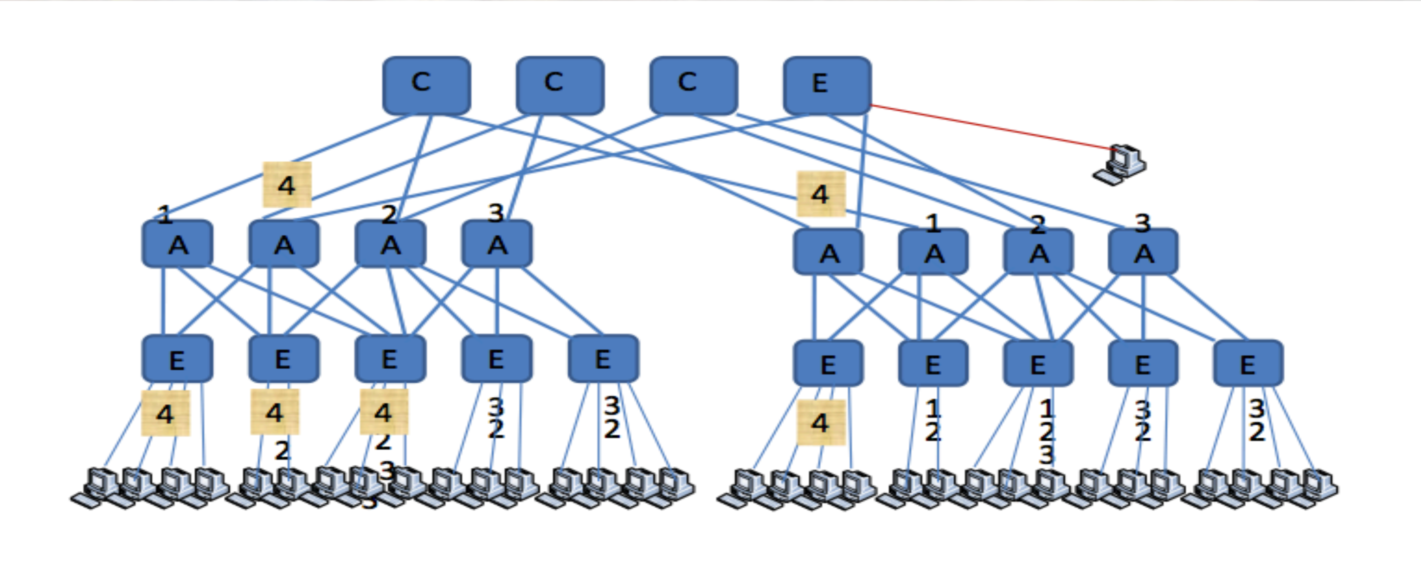
Types of Fat-tree Miswirings

- Phase 3: Reactivate isolated switches due to end-host misplacement
  - Host connected to an aggregation switch



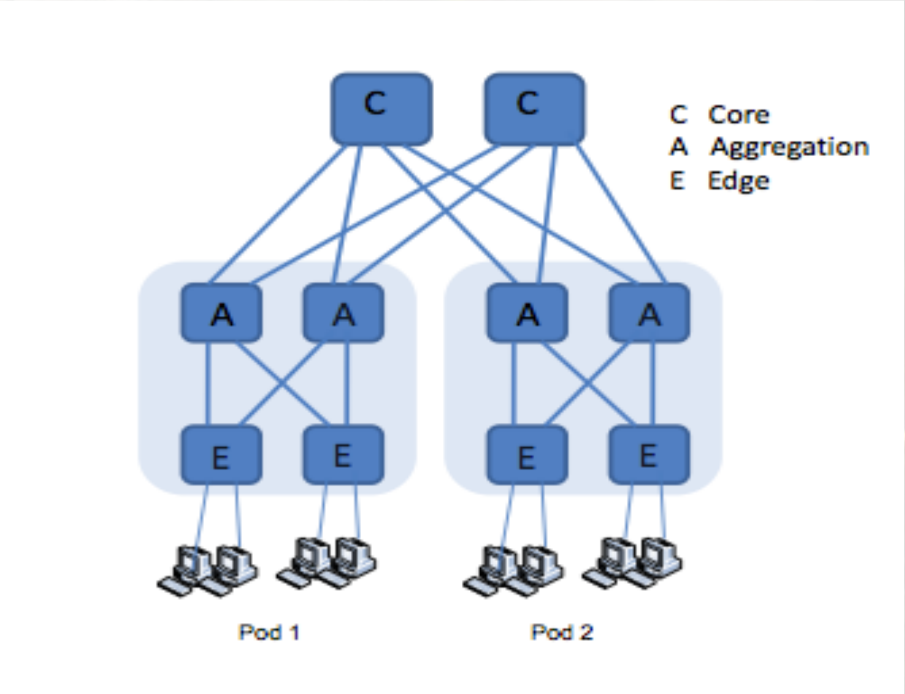
Types of Fat-tree Miswirings

- Phase 3: Reactivate isolated switches due to end-host misplacement
  - Host connected to a core switch



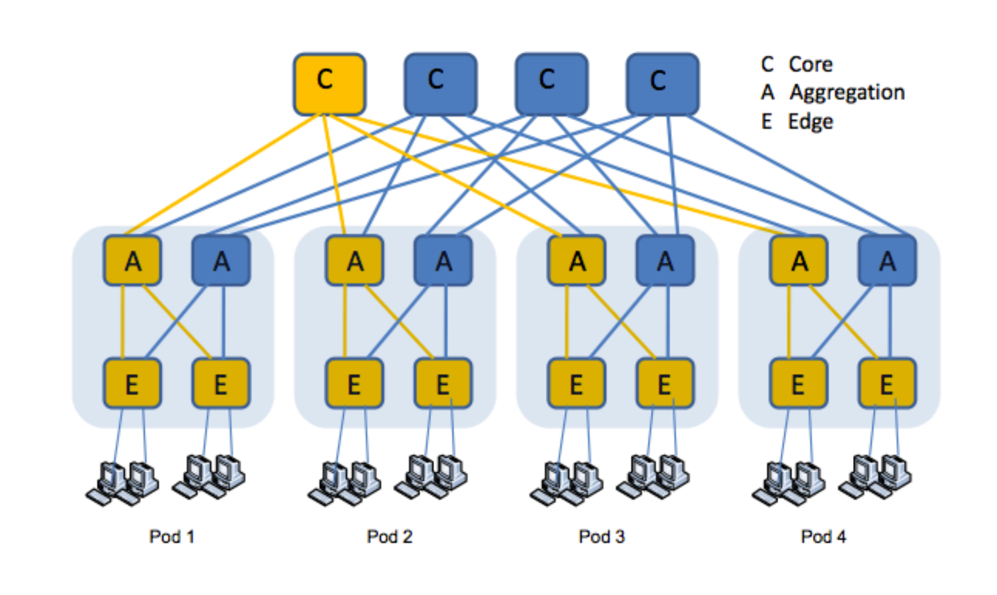
Types of Fat-tree Miswirings

- Phase 4: Multiple links from a core switch to different aggregation switches within a putative pod

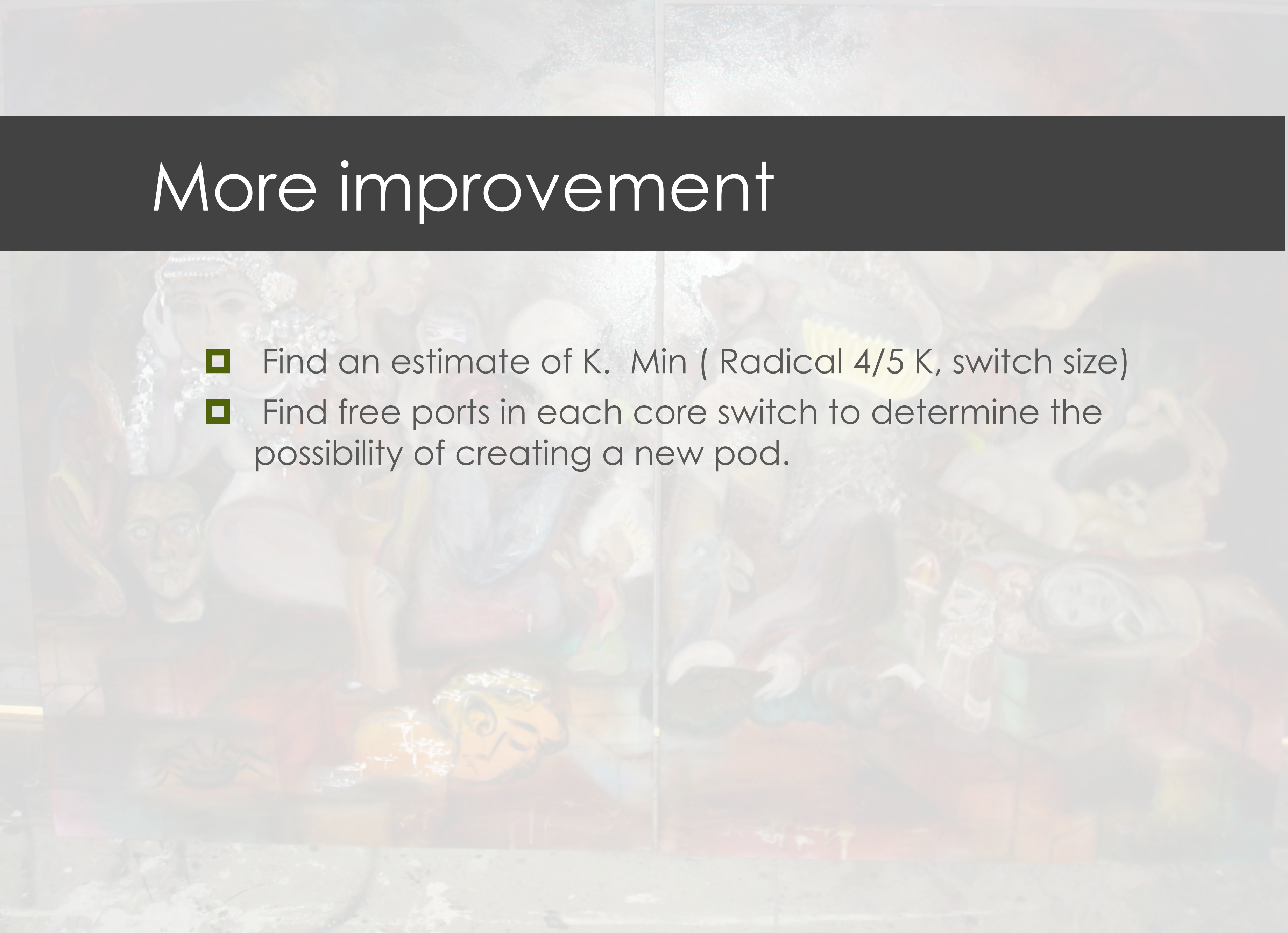


Improving Datacenter Performance

- Improve reachability and bandwidth within a pod:
  - Detection of Missing Links within a pod
  - No more than K/4 missing links can be detected.
- Improving reachability between pods:
  - Core switches building reachability single-root trees.
- This algorithm can be improved with given preferences.



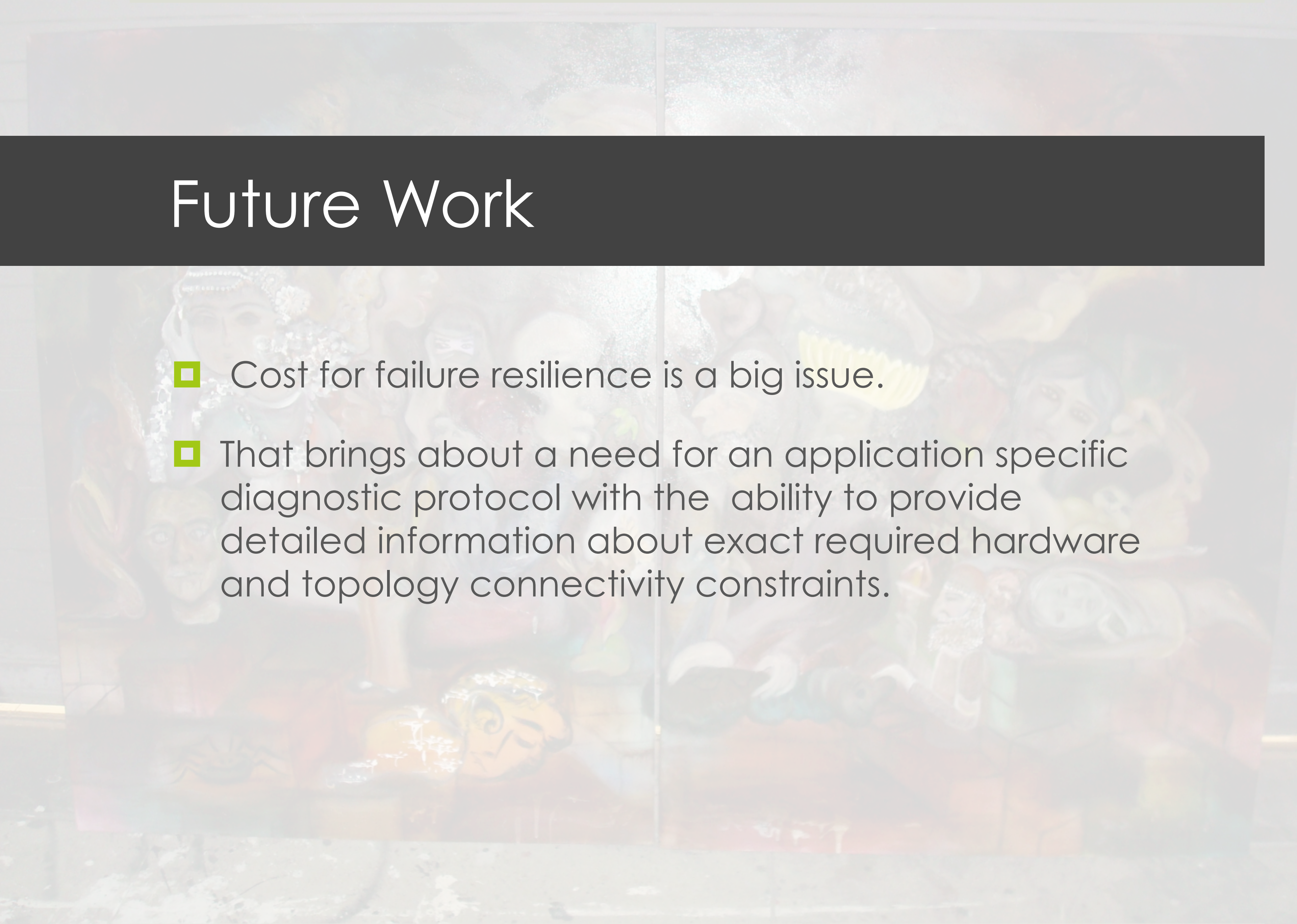
More improvement



- Find an estimate of K. Min (Radical 4/5 K, switch size)
- Find free ports in each core switch to determine the possibility of creating a new pod.

Summary

- Datacenter architecture has to be scalable.
- It is easy to make mistakes in datacenter wirings.
- 3-level Fat-tree is a scalable architecture.
- We developed a diagnostic protocol that detects badwirings based on fat-tree architecture.
- The protocol returns a prioritized list of recommended changes to improve performance.



Future Work

- Cost for failure resilience is a big issue.
- That brings about a need for an application specific diagnostic protocol with the ability to provide detailed information about exact required hardware and topology connectivity constraints.

Q&A