Clean Slate Design Approach to Networking Research

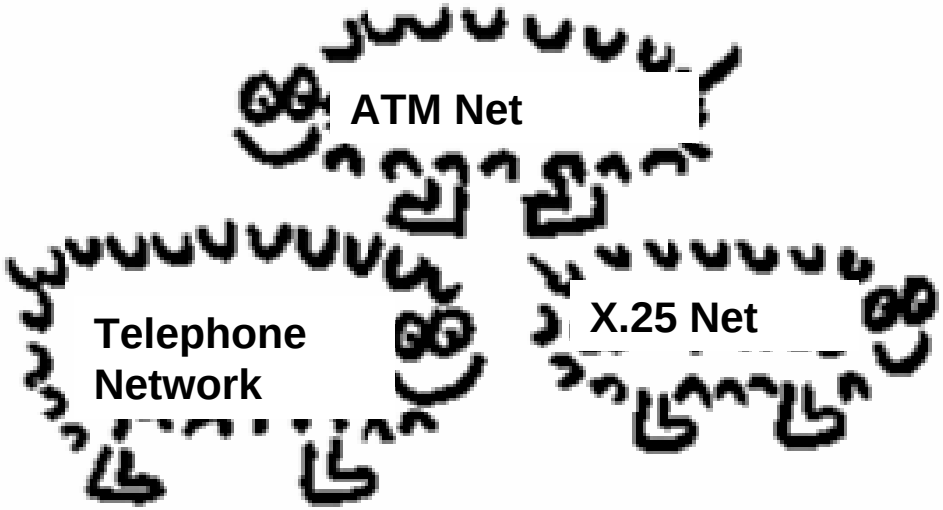
Hui Zhang

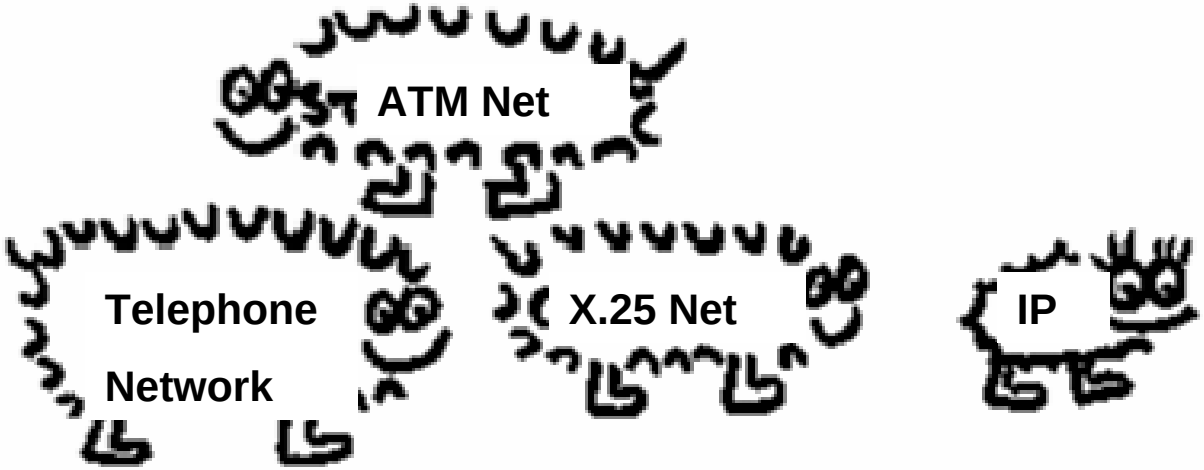
School of Computer Science

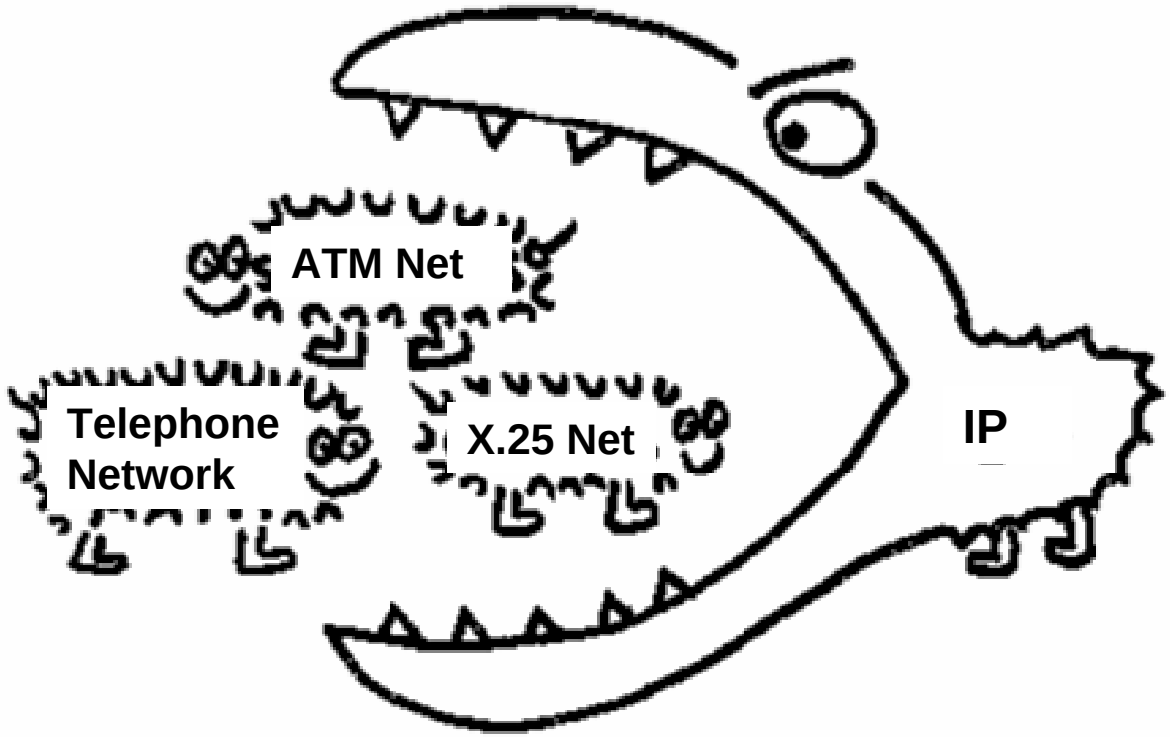
Carnegie Mellon University

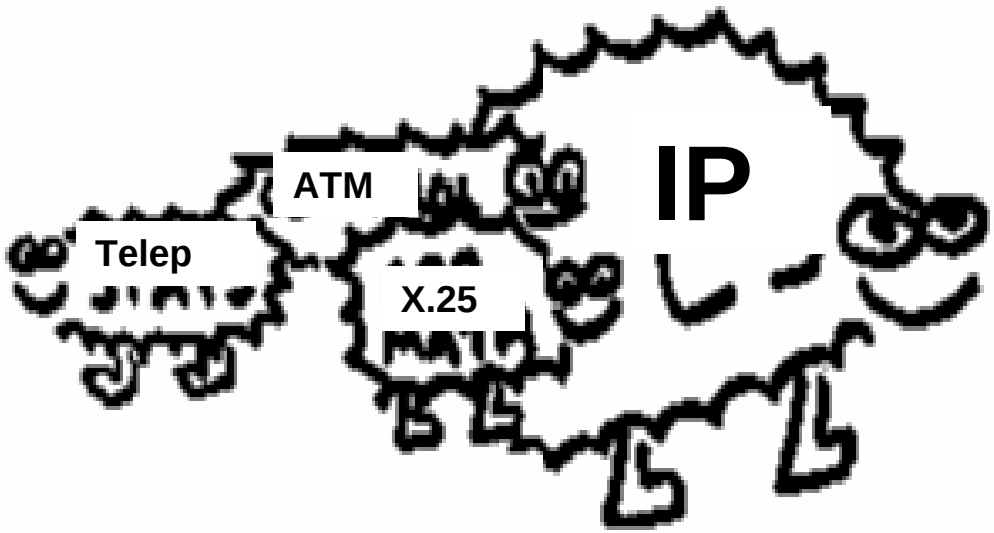
http://www.cs.cmu.edu/~hzhang

http://100x100network.org/


ATM Net
Telephone
Network
X.25 Net


ATM Net
Telephone
Network
X.25 Net
IP


ATM Net
Telephone
Network
X.25 Net
IP

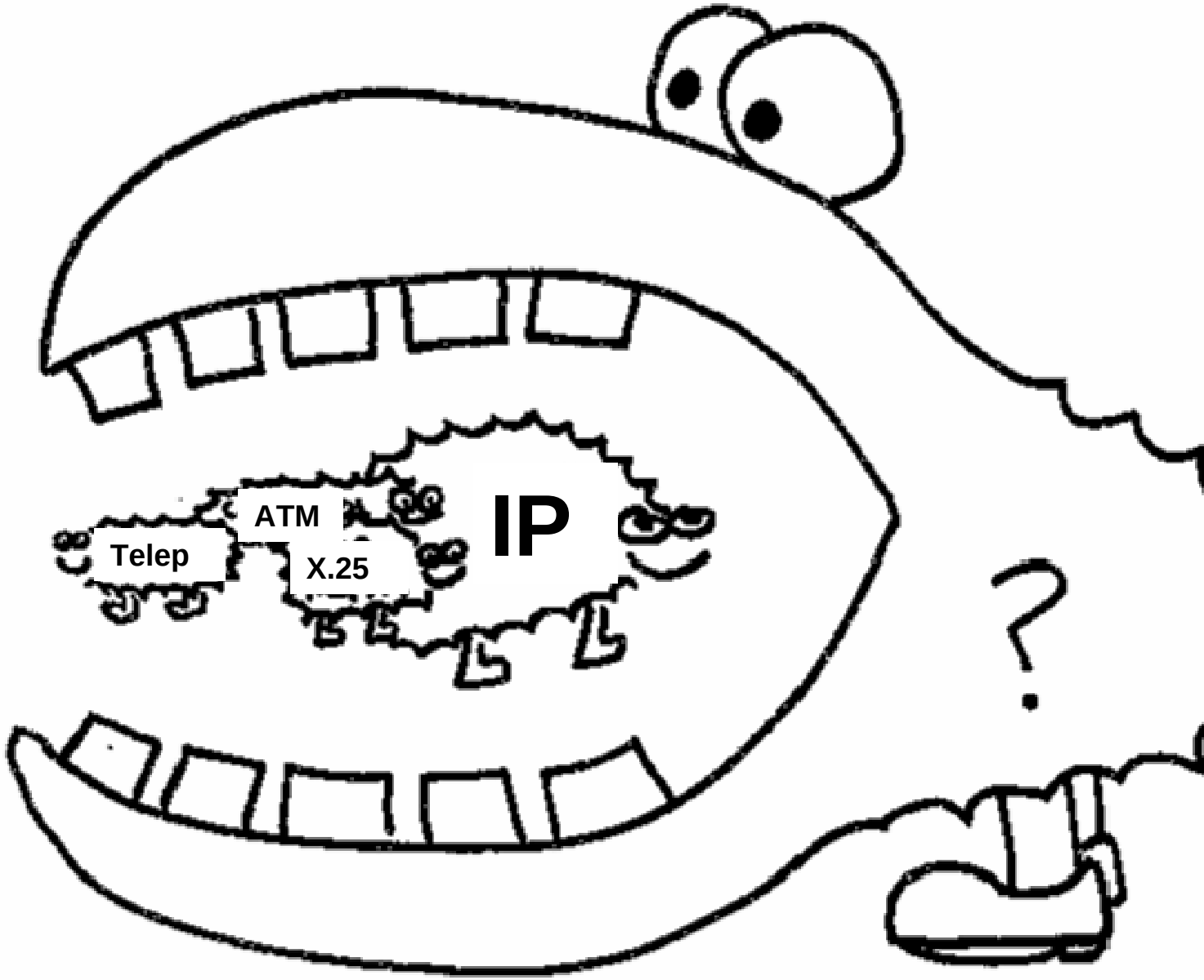

IP
ATM
Telep
X.25

The Success of the Internet and IP

- **The Internet**
  - Modest beginning with deep academic root
  - Global network with fundamental impact on society
- **IP was well suited for its pioneering role**
  - Global addressing scheme
  - Internetworking architecture
  - Best-effort reachability
- **Success is a double-edged sword → the world demands more from IP and the Internet**
  - Converged communication services
  - Dependability, privacy and security, economic sustainability

Networking Research

- **Internet and IP success is also a double-edge sword for researchers**
- **Research only on incremental fix to Internet and IP?**
  - IP and Internet are good enough
  - IP and Internet are difficult to change


ATM
IP
Telep
X.25
?

Clean Slate Design Approach to Networking Research

- **How would we design the network if we were to design it again from scratch?**
  - Not bound by existing design decisions
  - But take advantage the benefit of hindsight and the lessons we have learned

Clean Slate Project

- **Large Information Technoloy Research (ITR) project funded by National Science Foundation (NSF), starting November 2003**
- **Multiple institutions**
  - Carnegie Mellon University (lead institution), including Pittsburgh Supercomputing Center (PSC)
  - Fraser Research
  - Stanford
  - Berkeley
  - Rice
  - ATT Research
  - Internet 2

Clean Slate Project

- **100x100 means**
  - At least 100Mbps to 100 million households
  - 1 Gbps to 1 million small businesses
- **Exact numbers are not as important, but we would like to focus on a specific network**
  - Consider the network as a whole
  - Consider technology trends for scaling, cost-effectiveness, future-safeness
  - Architect with explicit considerations of economics, dependability, security
  - Design with explicit goals of enabling tractable analysis and modeling

Why Clean Slate Design?

- **A powerful research methodology that helps to crystallize the issues**
  - Smalltalk, Multics, Unix, TCP/IP
- **A concrete and complete different design point highlights possibilities**
- **Understanding the target first helps to plan the trajectory of evolution**

Why Clean Slate Design?

- **A mind set that may result in different research**
- **Incremental approach to security**
  - How to detect and stop Blaster, Code Red?
- **Clean slate design approach to security**
  - What would be the fundamental capability of a strategic adversary?
  - What are the fundamental limitations/possibilities of any network-based or host-based security mechanism?
  - What should be the minimal & necessary set of layer 3 security mechanism?

Research Directions

- Tradeoff between organic network growth vs. structured network design
- Large scale wireless and fiber access networks
- Load-balanced backbone networks
- End-to-end lossless flow control
- Economic informed network design
- Network forensics & disconnect-default communication model
- Network-wide control & management

A Clean Slate 4D Approach to Network Control and Management

Hui Zhang

Carnegie Mellon University

Joint work with

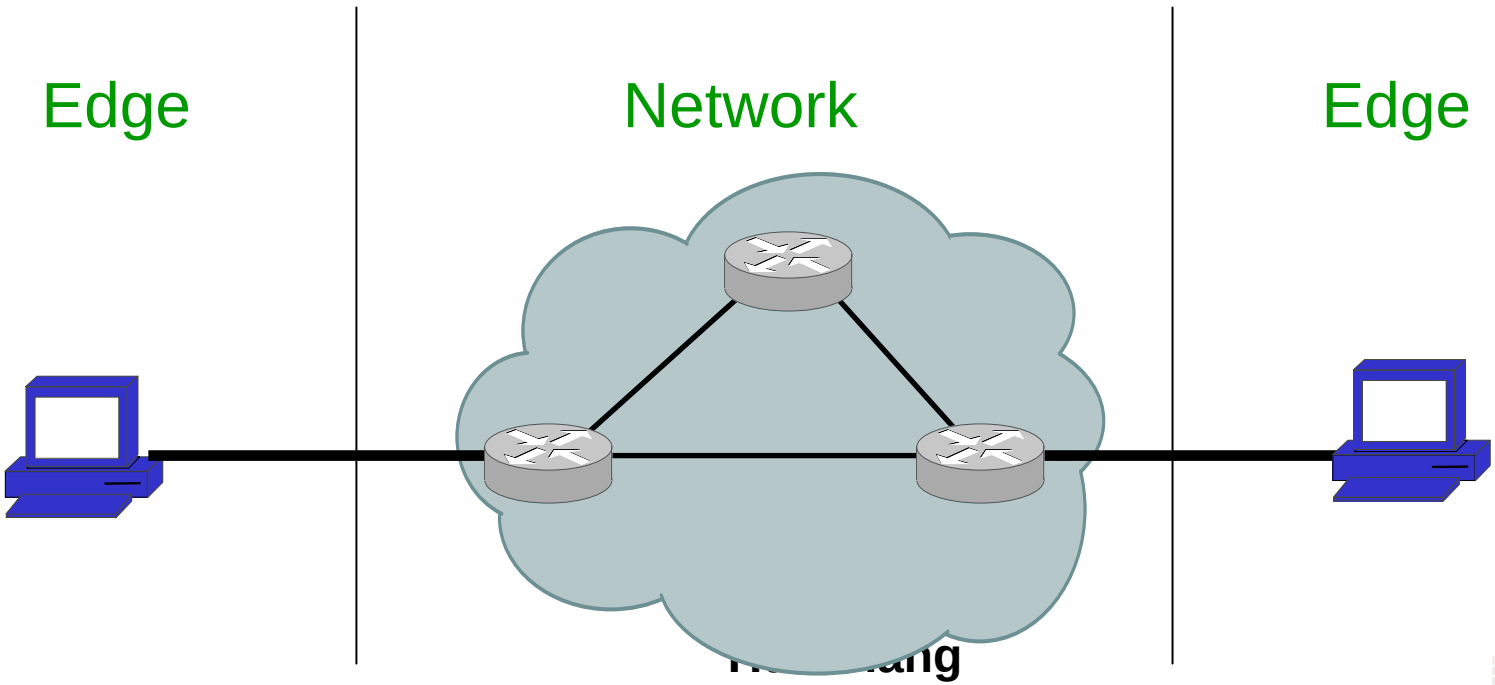
Albert Greenberg, Gisli Hjalmtysson

David Maltz, Andy Myers, Jennifer Rexford, Geoffrey Xie,

Hong Yan, Jibin Zhan

Stateless IP Architecture

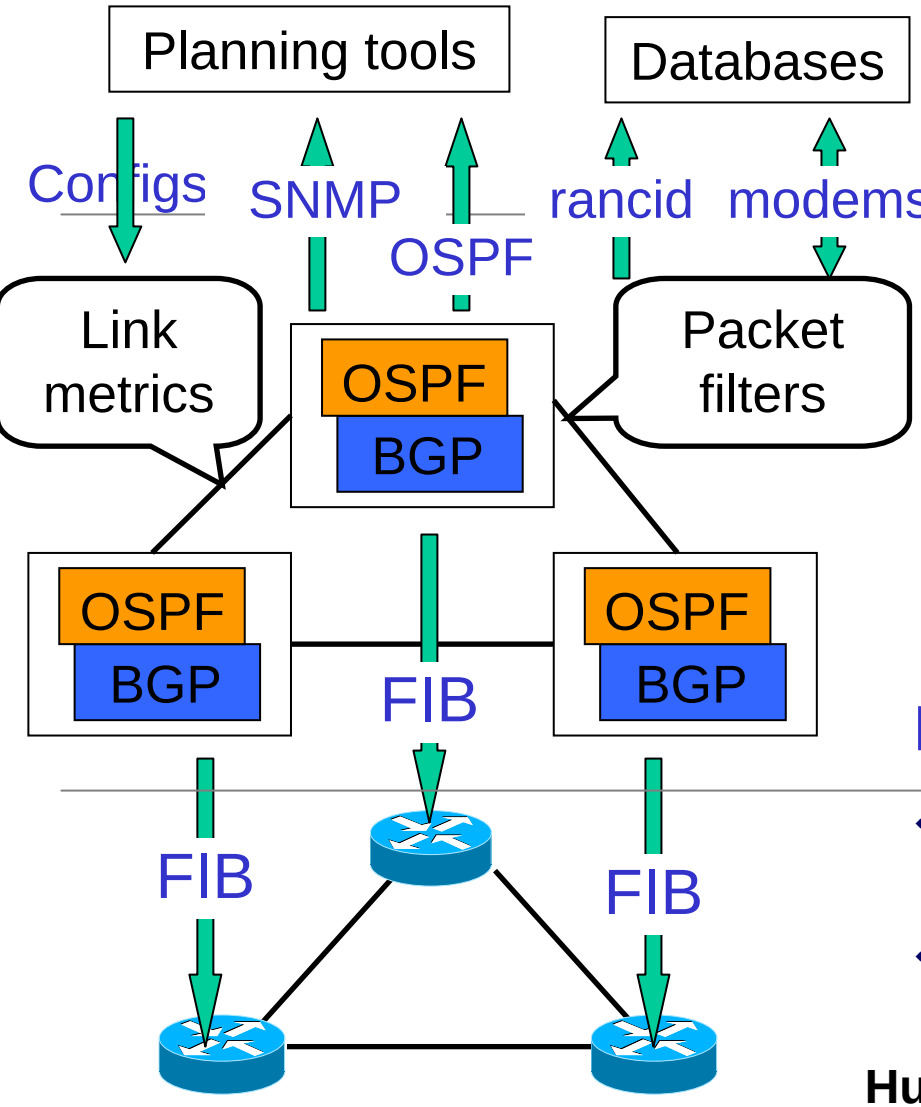
- **Smart hosts, dumb network**
- **Network moves IP packets between hosts**
- **Services implemented on hosts**
- **Keep state at the edges**



An Accident of History

Shell scripts

Tomography



Management Plane

- **Figure out what is happening in network**
- **Decide how to change it**

Control Plane

- **Multiple routing processes on each router**
- **Each router with different configuration program**
- **Huge number of control knobs: metrics, ACLs, policy**

Data Plane

- **Distributed routers forwarding packets**
- **FIBs, Access control, NAT, tunnels**

Hui Zhang

100 x 100

An Accident of History

Shell scr

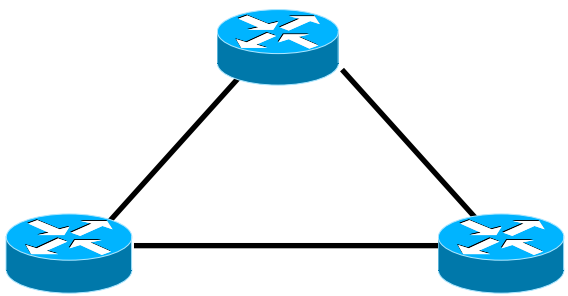
Tomography

Pl

atabases

OSPF

BGP



Management Plane

- Figure out what is happening in network
- Decide how to change it

l Plane

uting processes

nt

- **Based on FIB**

State everywhere!

- **Dynamic state** in FIBs
- **Configured state** in settings, policies, packet filters
- **Programmed state** in magic constants, timers
- Many dependencies between bits of state

State updated in uncoordinated, decentralized way!

Inside a Single Network

Shell scr

Traffic Eng

Management

- Figu

Pl

atabases

Con

State everywhere!

- **Dynamic state**
- **Configured**
- **Programm**
- Many d

Logic everywhere!

- **Path Computation** built into routing protocols
- **Routing Policy** distributed across the routers
- **Packet Filters** placed by tools in Mng. Plane

No way to arbitrate inconsistencies between logic

ated, decentralized way!

icies, packet filters

onstants, timers

n bits of state

Packet filters

- **Forwardi**
- **Based on FIB or**

Hui Zhang

A Study of Operational Production Networks

- **How complicated/simple are real control planes?**
  - What is the structure of the distributed system?
- **Use *reverse-engineering* methodology**
  - There are few or no documents
  - The ones that exist are out-of-date
- **Anonymized configuration files for 31 active networks (>8,000 configuration files)**
  - 6 Tier-1 and Tier-2 Internet backbone networks
  - 25 enterprise networks
  - Sizes between 10 and 1,200 routers
  - 4 enterprise networks significantly larger than the backbone networks

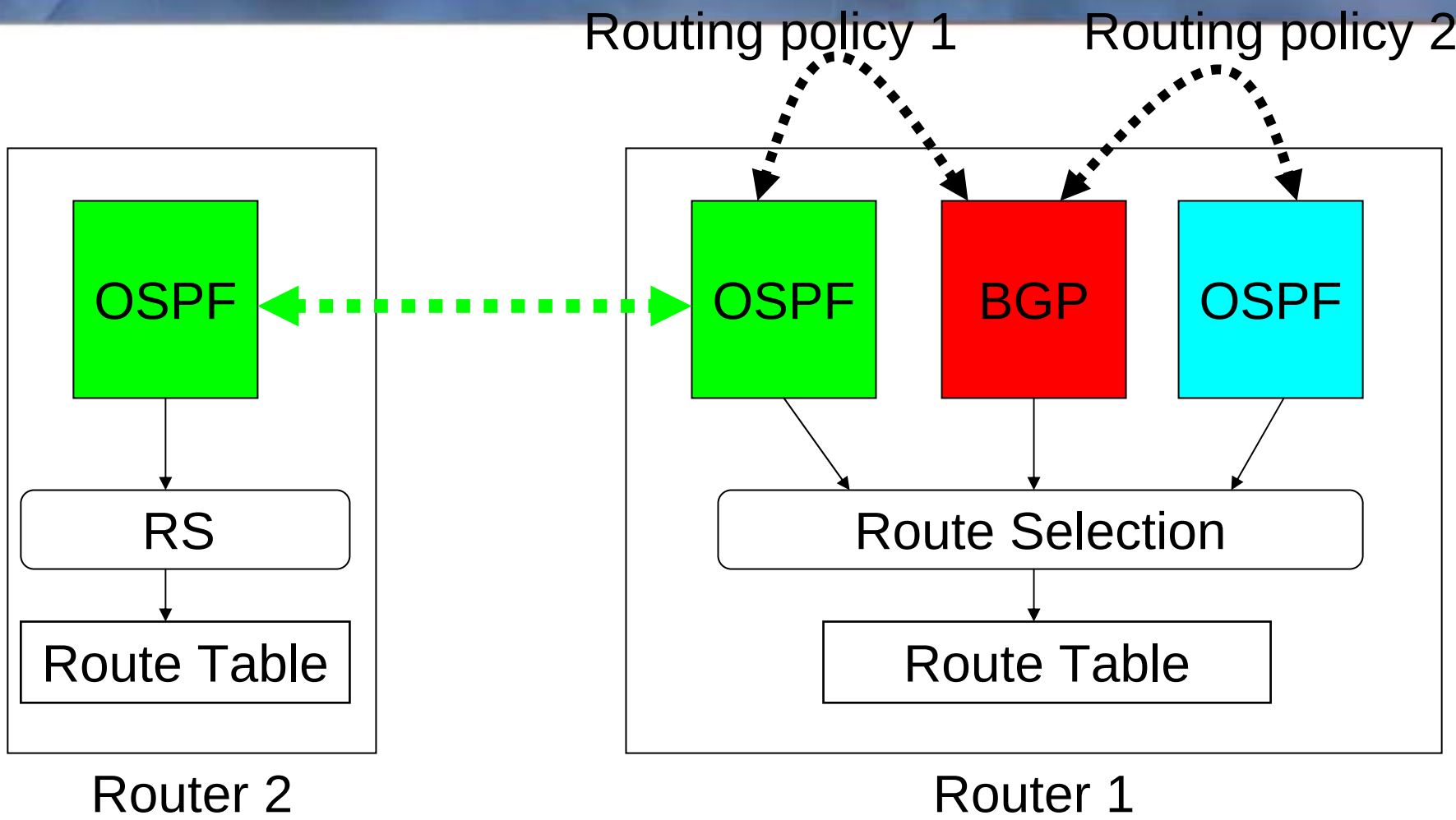
Router Configuration Files

- **interface Ethernet0**
- ** ip address 6.2.5.14 255.255.255.128**
- **interface Serial1/0.5 point-to-point**
- ** ip address 6.2.2.85 255.255.255.252**
- ** ip access-group 143 in**
- ** frame-relay interface-dlci 28**
- **router ospf 64**
- ** redistribute connected subnets**
- ** redistribute bgp 64780 metric 1 subnets**
- ** network 66.251.75.128 0.0.0.127 area 0**
- **router bgp 64780**
- ** redistribute ospf 64 match route-map 8aTzlvBrbaW**
- ** neighbor 66.253.160.68 remote-as 12762**
- ** neighbor 66.253.160.68 distribute-list 4 in**

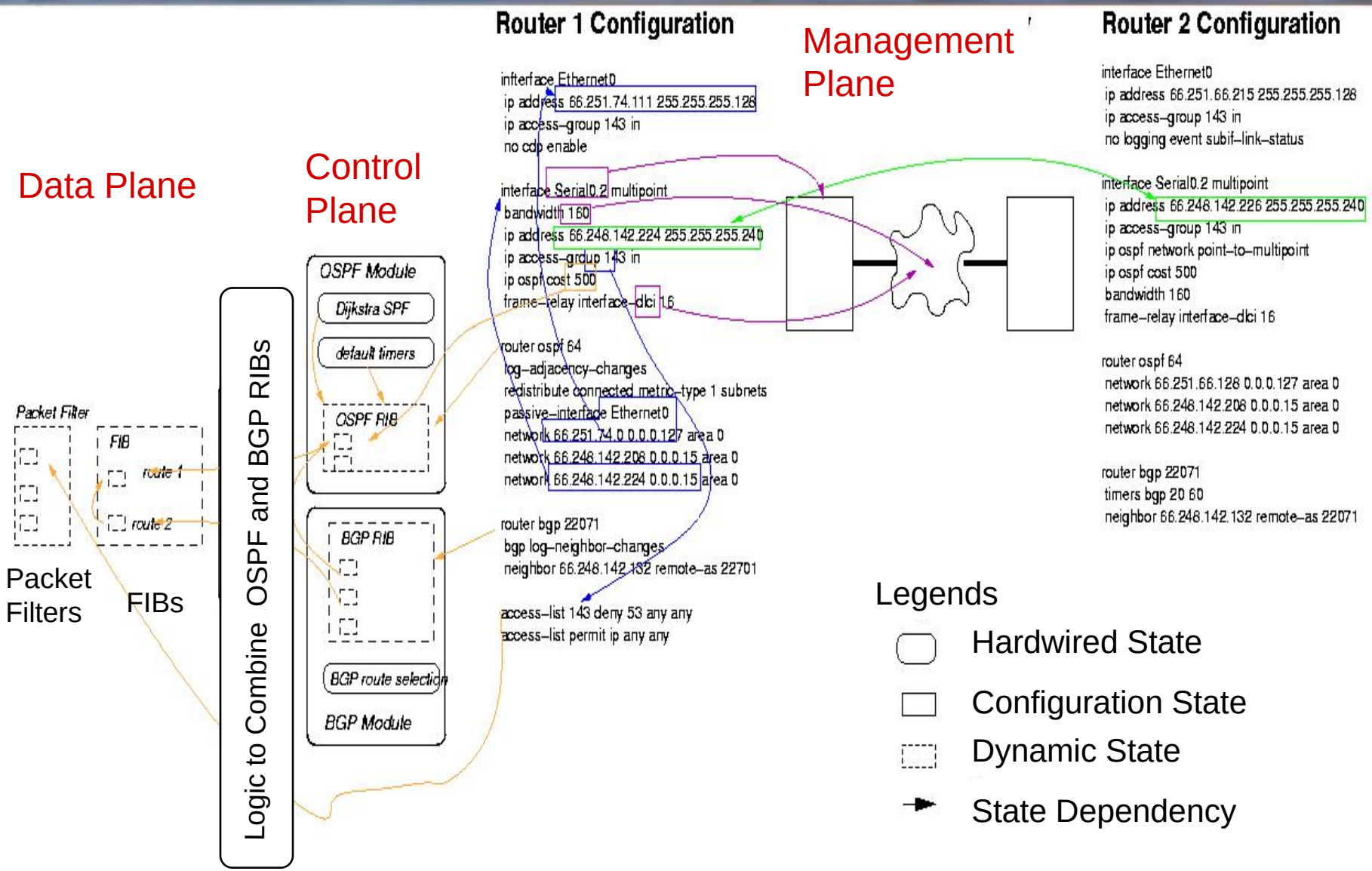
```
access-list 143 deny 1.1.0.0/16
access-list 143 permit any
route-map 8aTzlvBrbaW deny 10
 match ip address 4
route-map 8aTzlvBrbaW permit 20
 match ip address 7
ip route 10.2.2.1/16 10.2.1.7
```

Hui Zhang

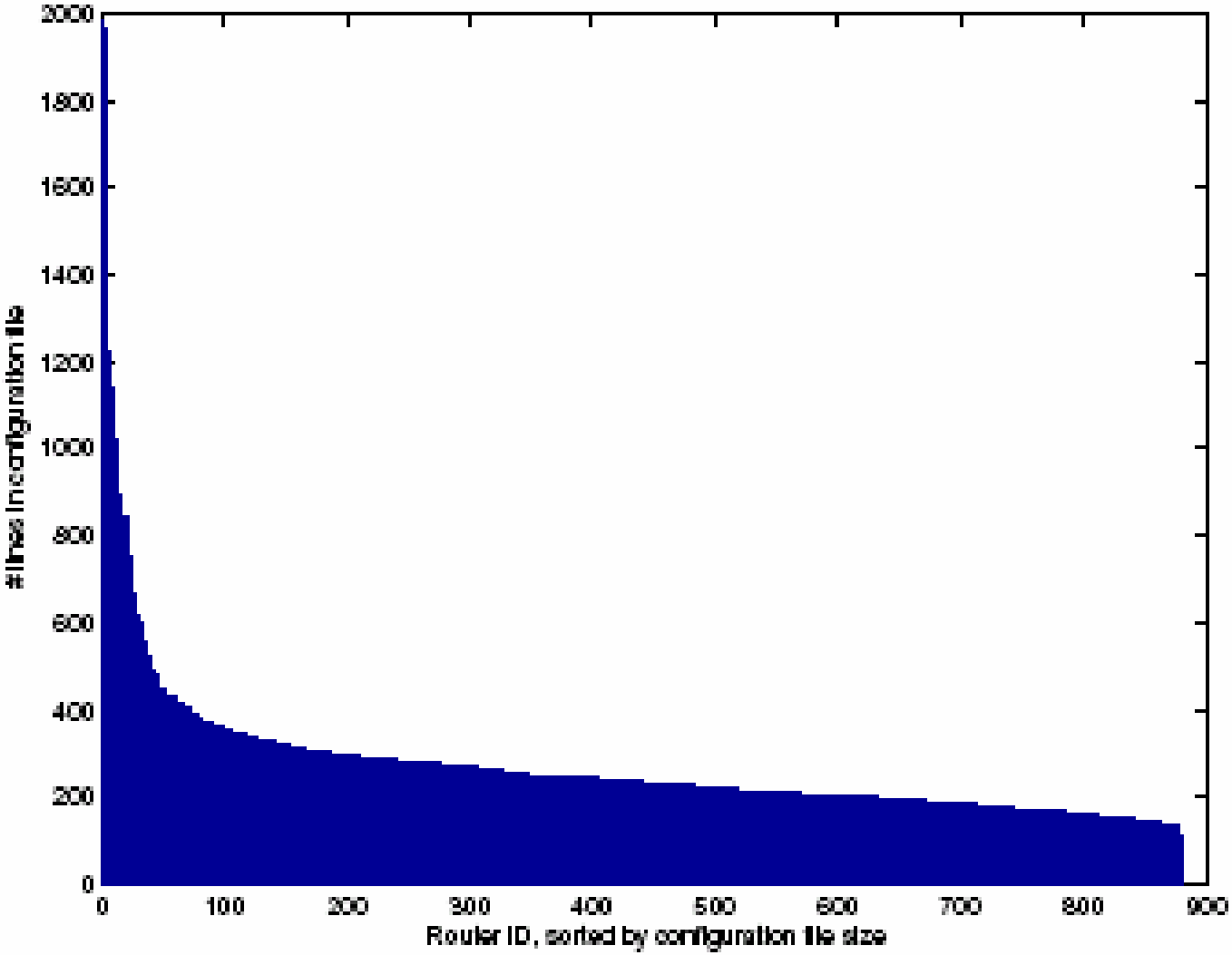
Routing Protocol Interactions



Complex Interaction of States



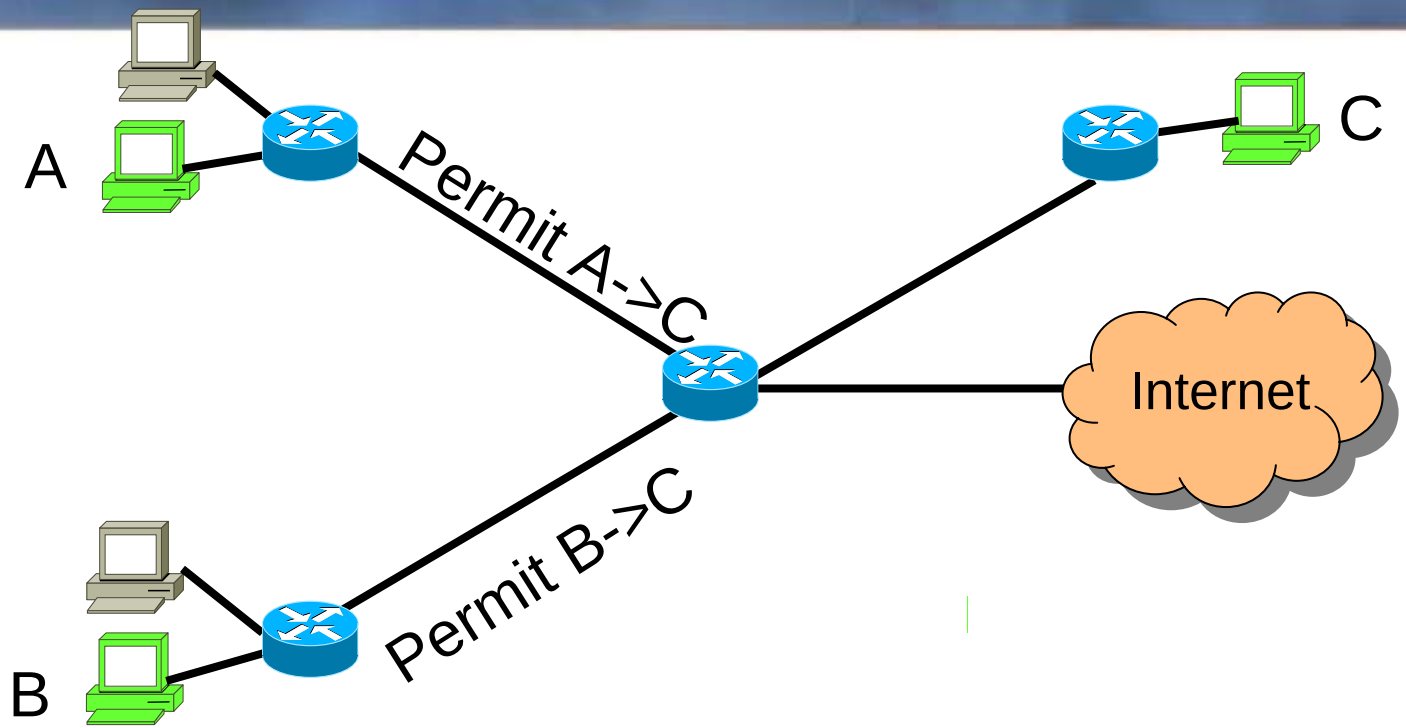
Configuration State for One Network



Hui Zhang

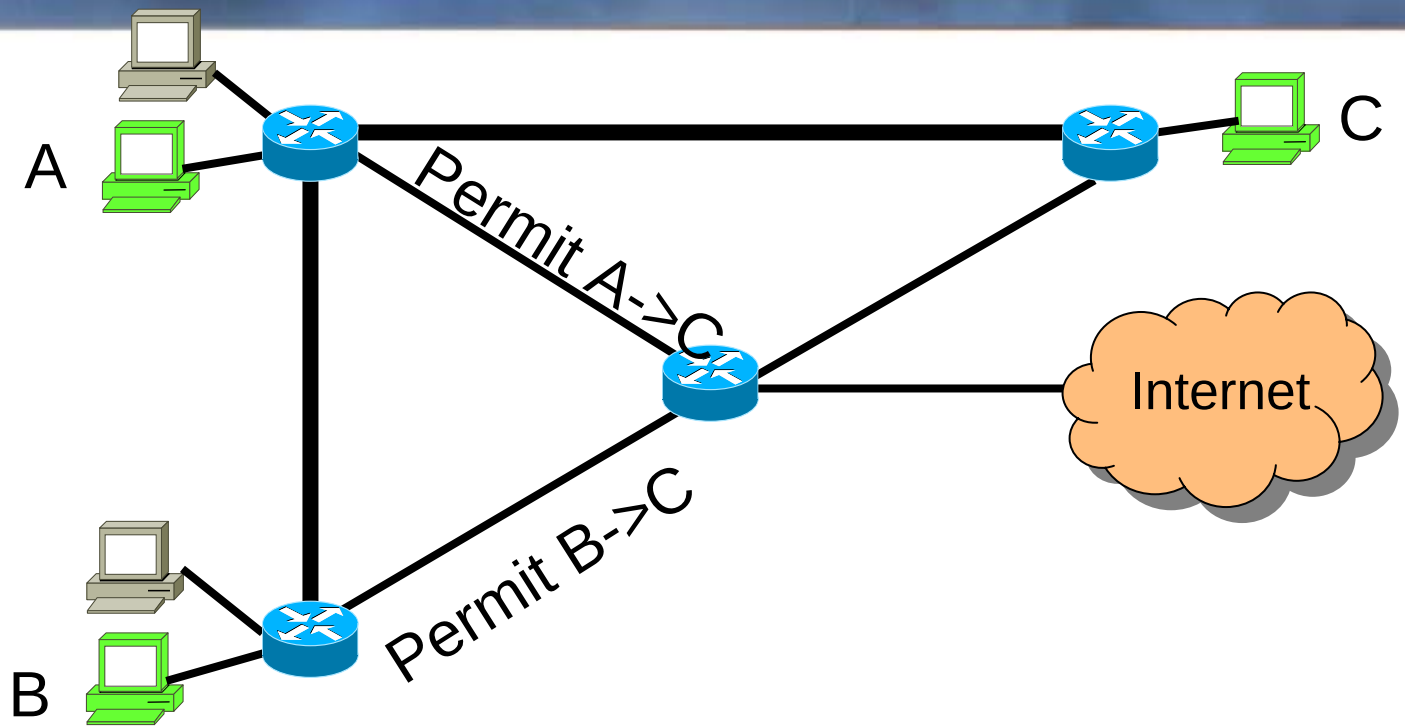
100 x 100

Reachability Example



- Enterprise with two remote offices
- Only A&B should be able to talk to server C

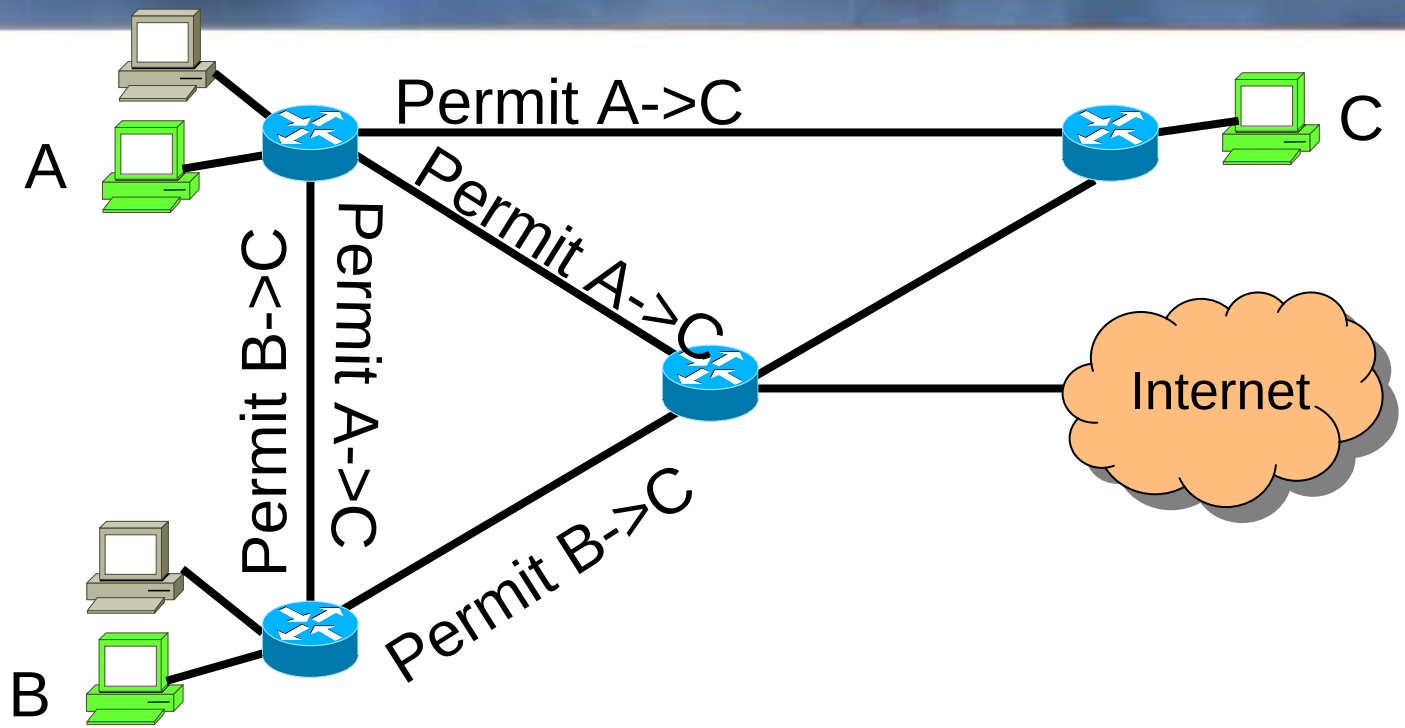
Reachability Example



- Network designers add two links for robustness
- Configure routing protocols to use new links in failure

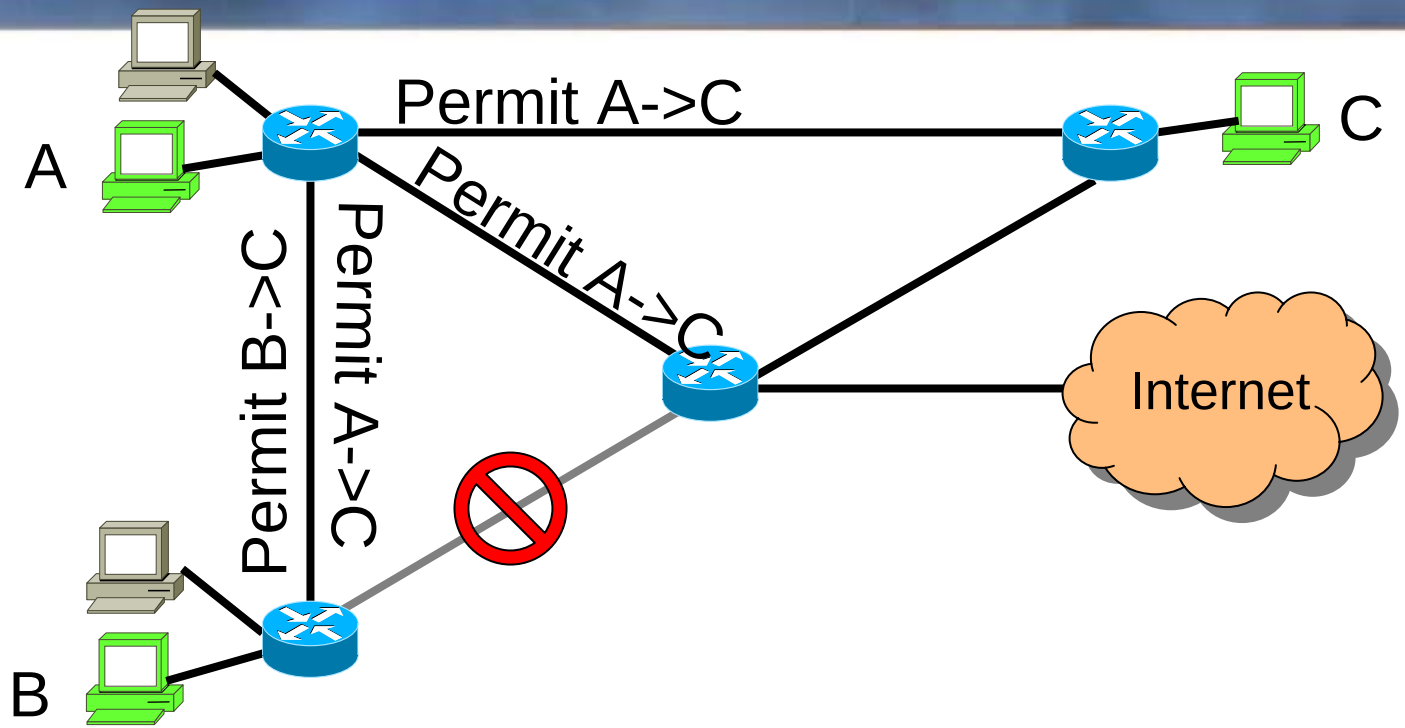
Hui Zhang

Reachability Example

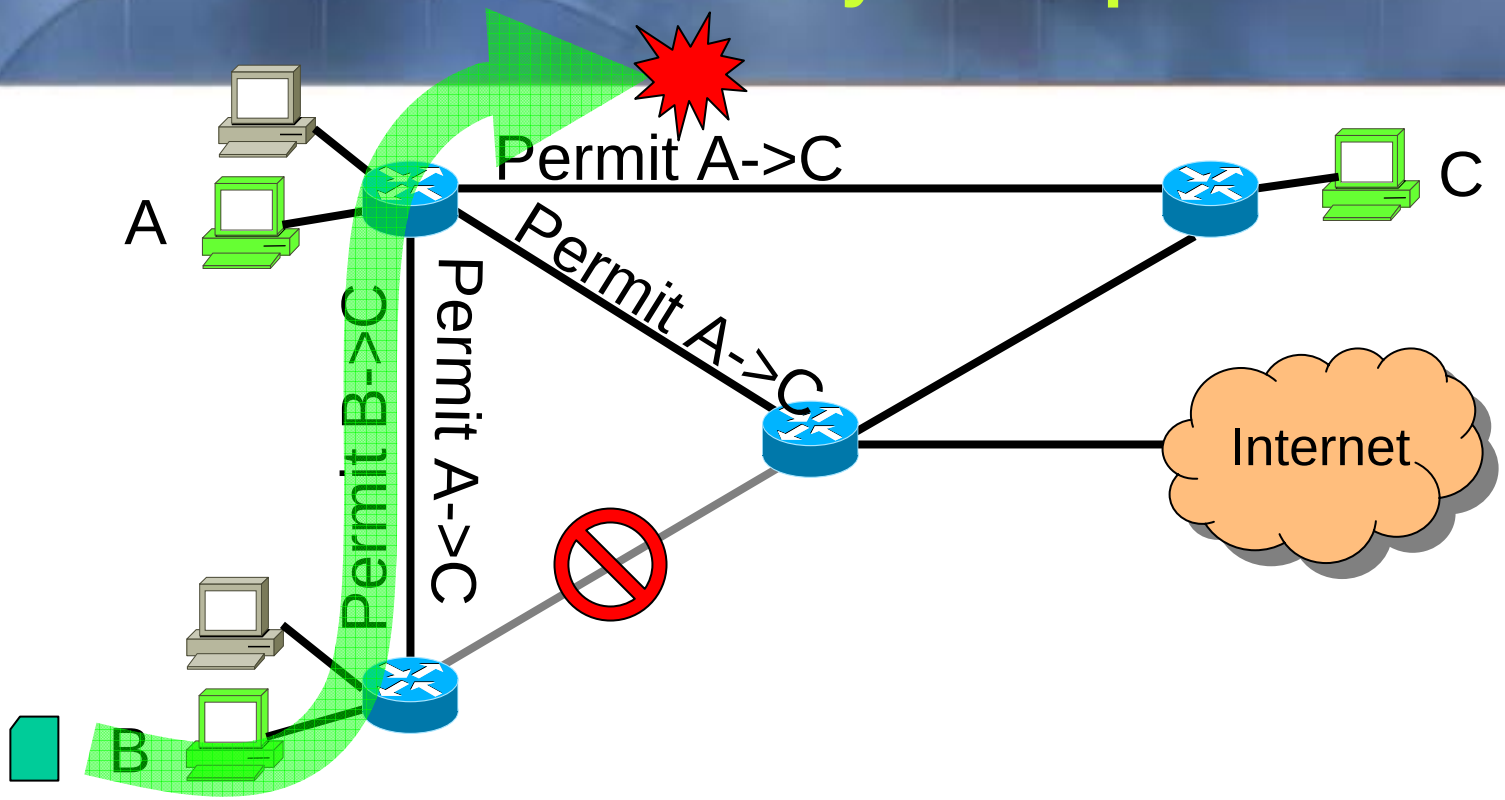


- **Designers apply packet filters to new links**

Reachability Example



Reachability Example

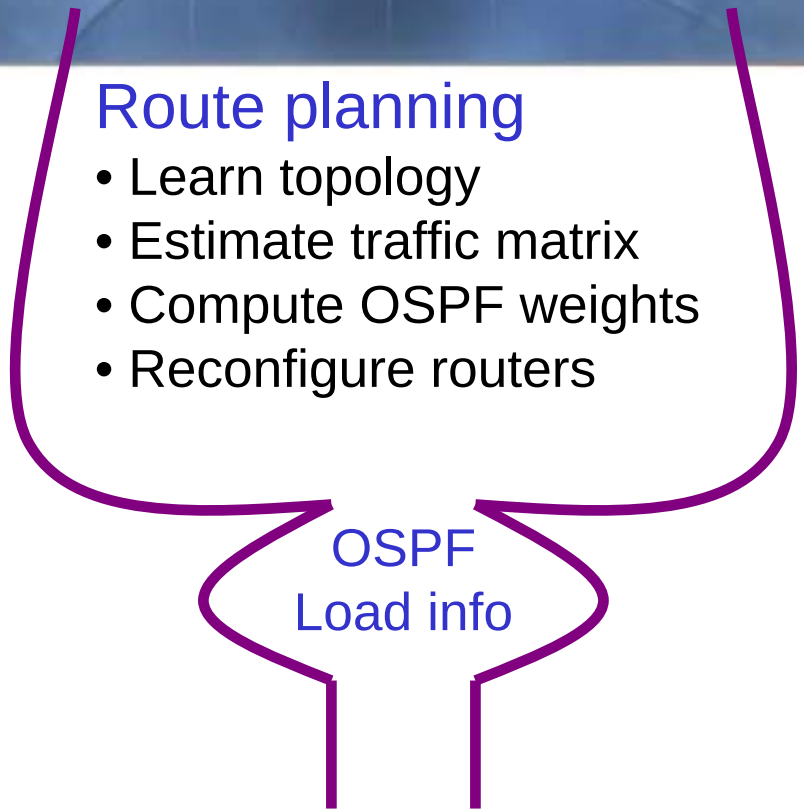


- **Packet from B->C dropped!**
- **Testing under normal conditions won't find this error!**

Need for Network-wide Control and Management

- **Higher level specification of network wide goals,**
  - Reachability matrix vs. per interface access control list
- **Dynamic coordination among diverse mechanisms:**
  - forwarding and access control
  - BGP route withdraw and access control list install

Another Example – Traffic Engineering



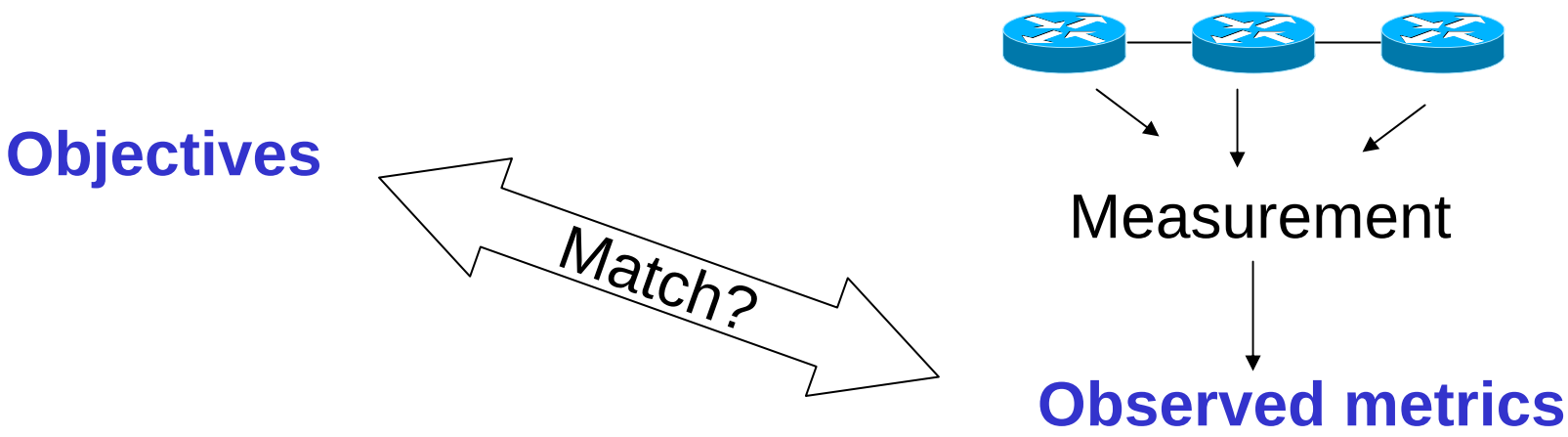
Management Plane

Control Plane

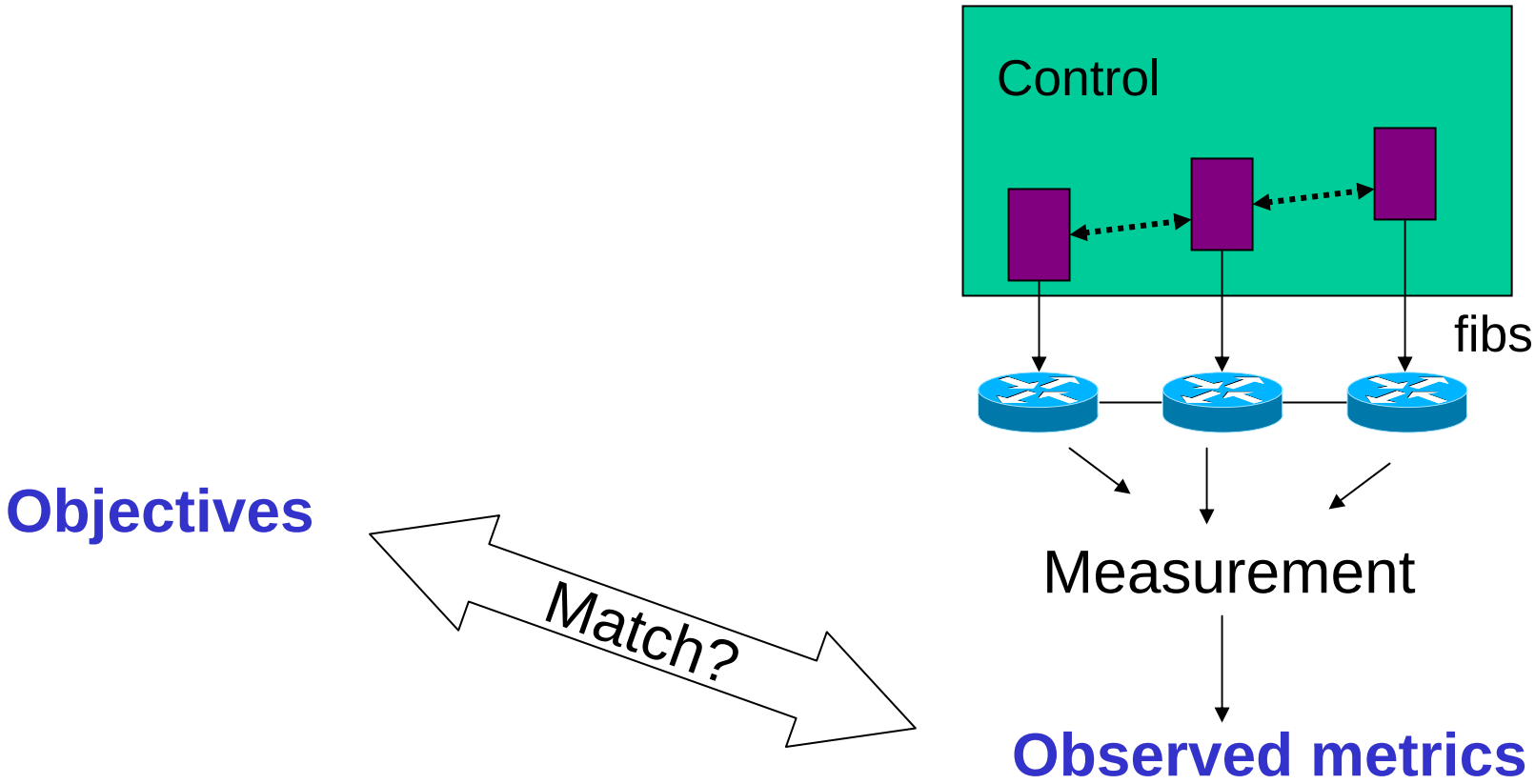
Data Plane

- **Must predict & undo effects of control plane**
- **Must translate solution into settings of control plane knobs**
- **Need ability to express desired solution**

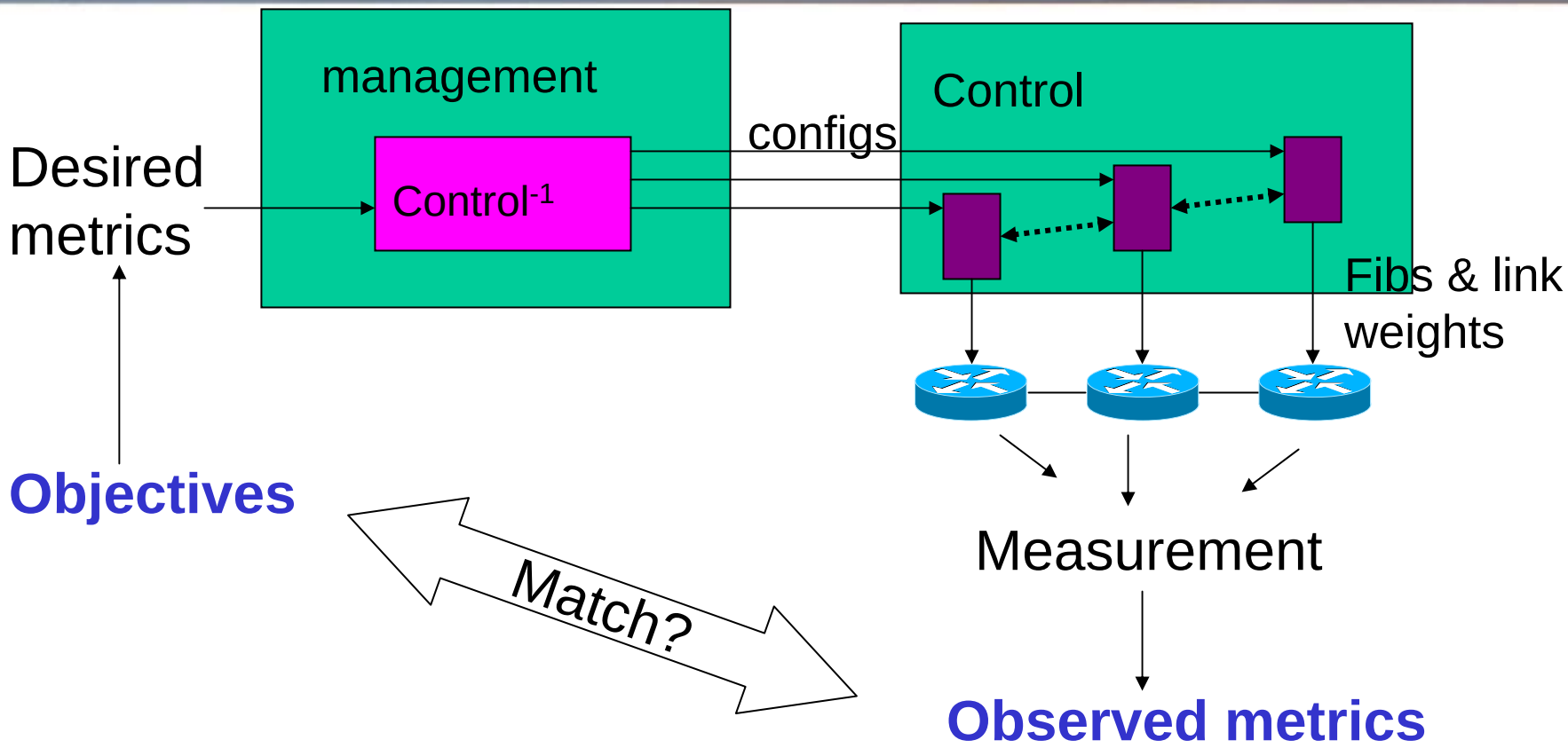
Indirect Expression of Goals



Indirect Expression of Goals



Indirect Expression of Goals



Systems of Systems

- **Systems are designed as components to be used in larger systems in different contexts, for different purposes, interacting with different components**
  - Example: OSPF and BGP are complex systems in its own right, they are components in a routing system of a network, interacting with each other and packet filters, interacting with management tools …
- **Complex configuration to enable flexibility**
  - The glue has tremendous impact on network performance
  - No high-level abstraction, no support for real-time coordination
  - State of art: multiple interactive distributed programs written in assembly language
- **Lack of intellectual framework to understand global behavior**

Key Challenge is Complexity

- **Too much focus on data plane and performance**
  - Encapsulation, congestion control, scheduling
- **Yet, the network is about coordination: control and management planes**
  - Distributed state management
  - Consequence of failing in control/management is severe
- **Status quo of control and management: extreme complex, non-linear, fragile, difficult to understand**

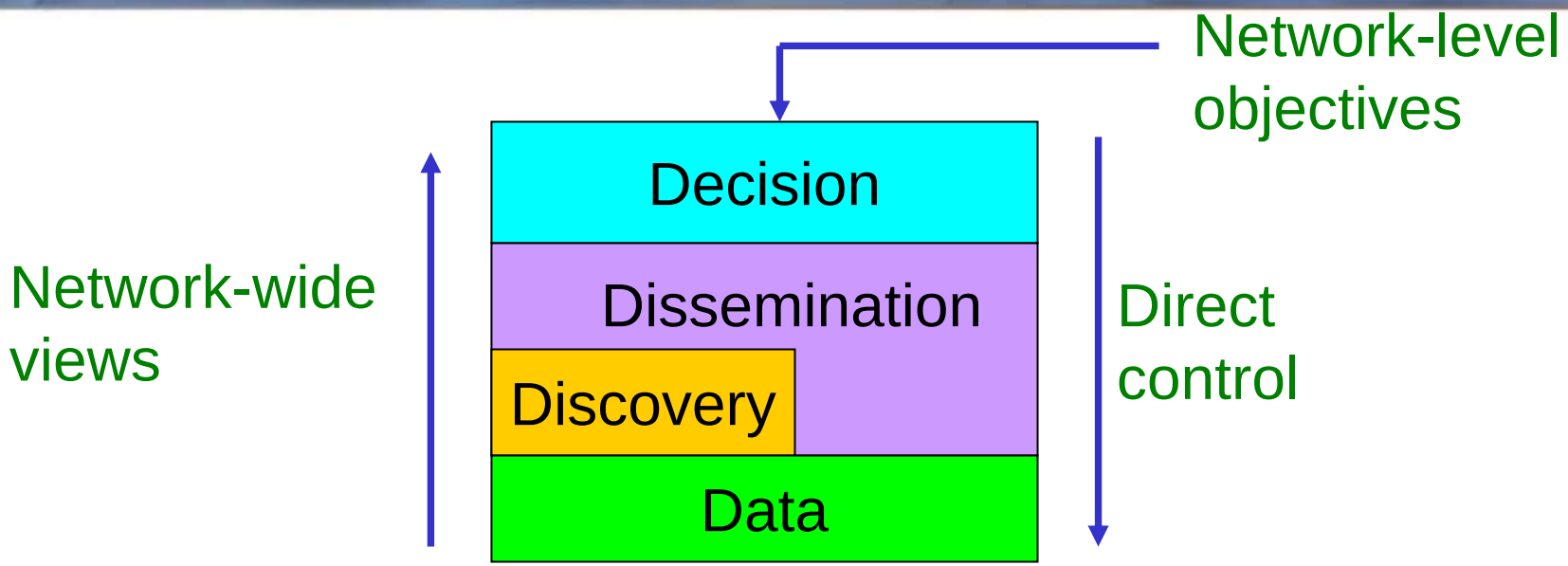
Are We Going to The Right Direction?

- **IP Control Plane function overloading**
  - Reachability
  - Policy control
  - Resiliency and survivability
  - Traffic Engineering, load balancing
  - VPN
- **Ethernet control plane overloading**
  - Spanning Tree, RSP, MSTP, vLAN, …
- **Complexity works against robustness, dependability, security**

Refactoring Control and Management Functions

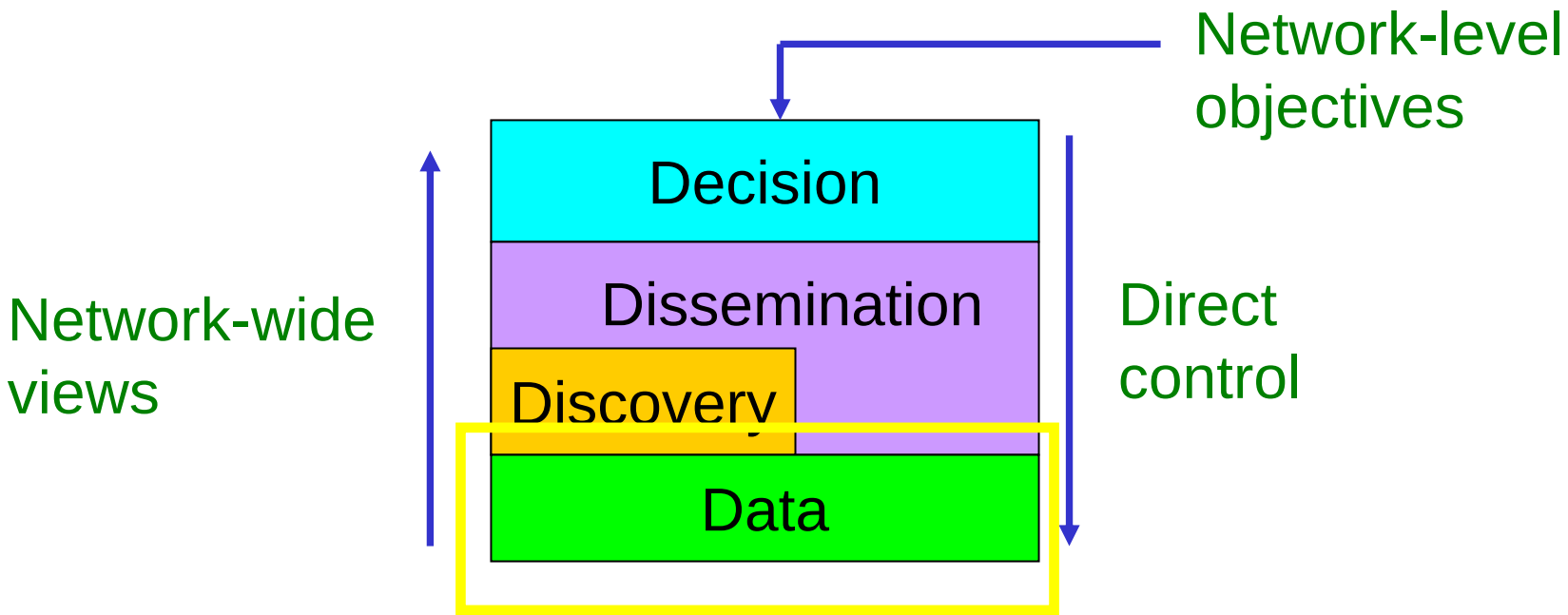
- **What's the right partitioning of functionality?**
- **What's the right abstractions?**
  - Good abstractions reduces complexity

Overview of the 4D Architecture



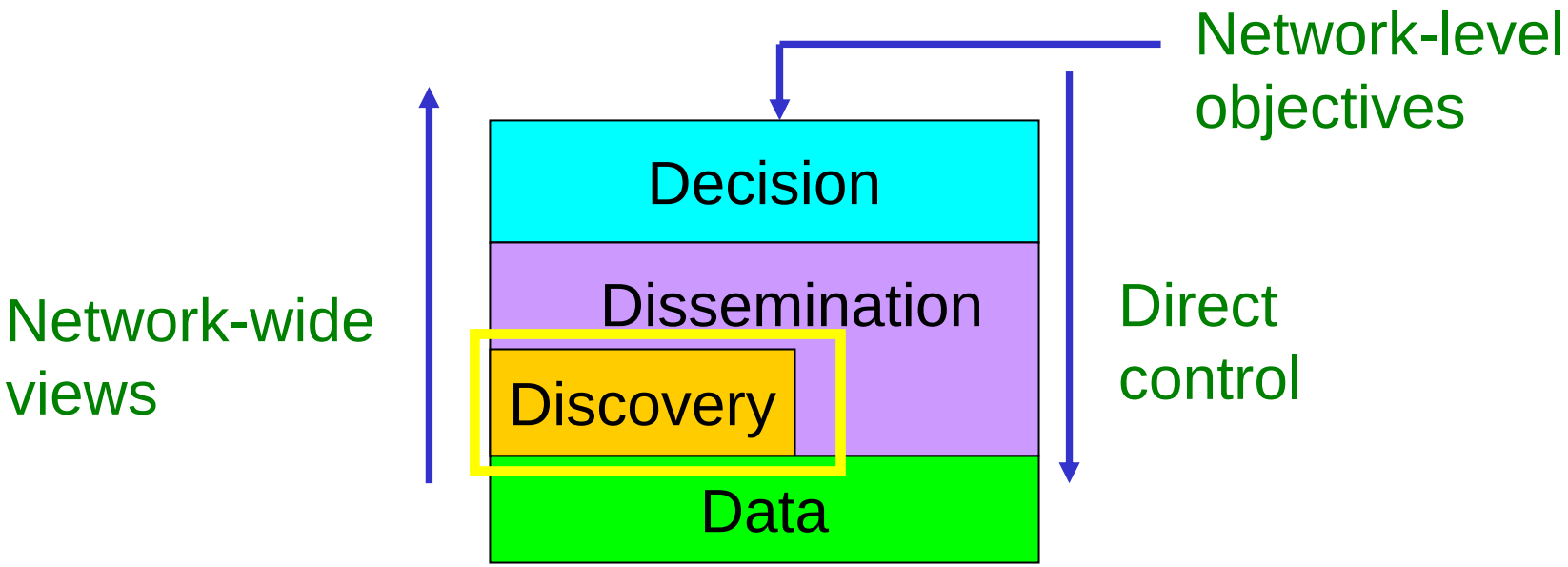
- **Centralized/replicated Decision Elements implement *all* decisions logic**
- **Decision Elements use views to compute data plane state that meets objectives, then directly writes this state to routers**

Overview of the 4D Architecture



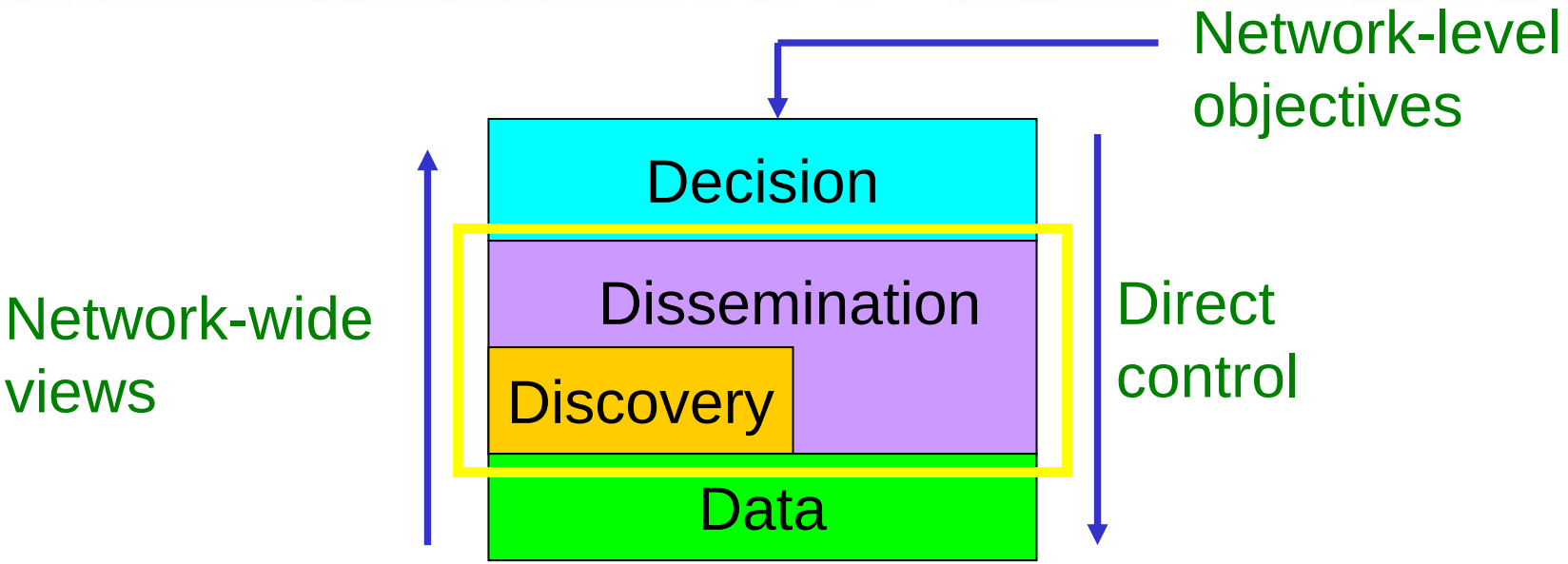
- *Data Plane:*
  - Modeled as set of distributed tables

Overview of the 4D Architecture



- *Discovery Plane:*
  - Each router discovers its own resources and its local environment

Overview of the 4D Architecture



- *Dissemination Plane:*
  - Provides a robust communication channel to each router
  - May run over same links as user data, but logically separate and independently controlled

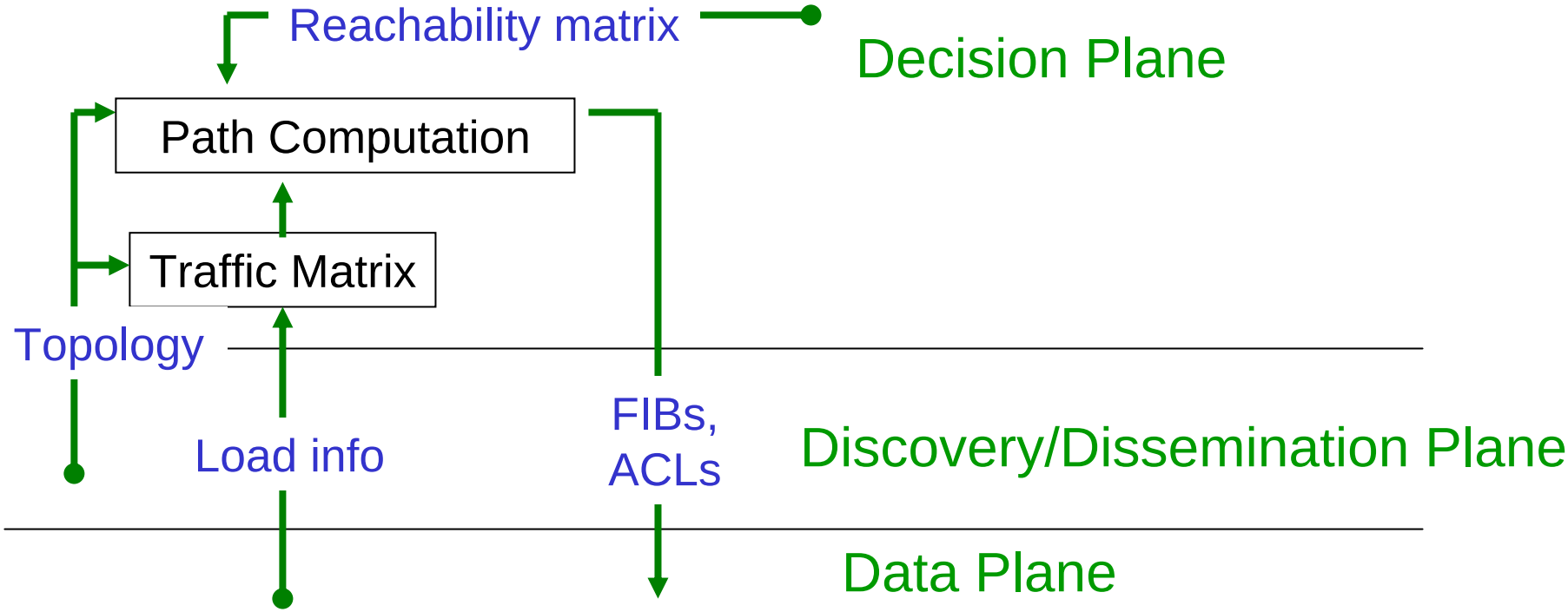
Devil's in the Detail

- **What are the identifiers? What are the scopes and persistence?**
  - E.g. interface card associated with hardware port, layer-two logical port, index for SNMP
    - What identifiers should be used for traffic statistics, hardware failure rates?
    - Should they survive reboots, replacement of interfaces?
  - Router identification
    - IP address? Router ID?
    - How to auto-configure?
      - Today: Addresses have to configured before a router can start communication

Simple Questions

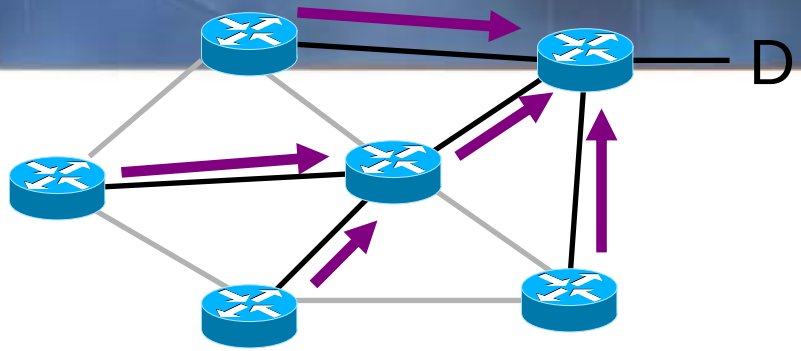
- **Should switches/routers be in the same address space as end hosts?**
  - End hosts hack into routers?
- **Communication channel for control and management**
  - Operational when data channel

Example – 4D Approach to Reachability Control

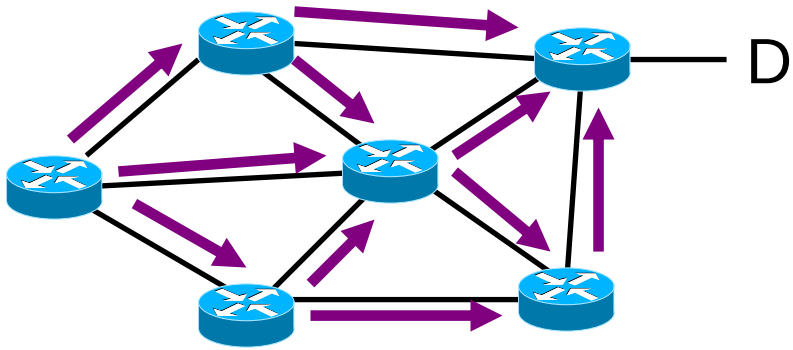


- **Reachability matrix** *directly expresses* **intended goal**
- **Path computation can** *jointly* **balance load and obey reachability constraints**
- **Packet filters installed only where needed, and changed when routing changes**

4D Enables Simpler and Better Traffic Engineering



- **OSPF normally calculates a single path to each destination D**
- **OSPF allows load-balancing only for equal-cost paths to avoid loops**
- **Using ECMP requires careful engineering of link weights**



Decision Plane with network-wide view can do more sophisticated optimization

Hui Zhang

4D Separates Distributed Computing Issues from Networking Issues

- **Distributed computing issues: protocols and network architecture**
  - Overhead
  - Resiliency
  - Scalability
- **Networking issues: decision logic**
  - Traffic engineering and service provisioning
  - Egress point selection
  - Tunnel management
  - Reachability control (VPNs)
  - Precomputation of backup paths

One Size Fits All?

- **Many different network environments**
  - Data center networks, enterprise/campus
  - Access, backbone networks
- **Many different forwarding**
  - Longest-prefix routing, exact-match switching, label switching
  - IP, MPLS, ATM, optical circuits
- **Many different objectives**
  - Routing, reachability, transit, traffic engineering, robustness

- **Today**
  - Different set of protocols for different data planes
    - STP for Ethernet
    - PNNI for ATM
    - OSPF/BGP for IP
  - Same protocols (logic) for different environments
    - Data center, campus, ISP
- **4D**
  - Common discovery & dissemination infrastructure
  - Customizable decision plane

Hui Zhang

100 x 100

The Feasibility of the 4D Architecture

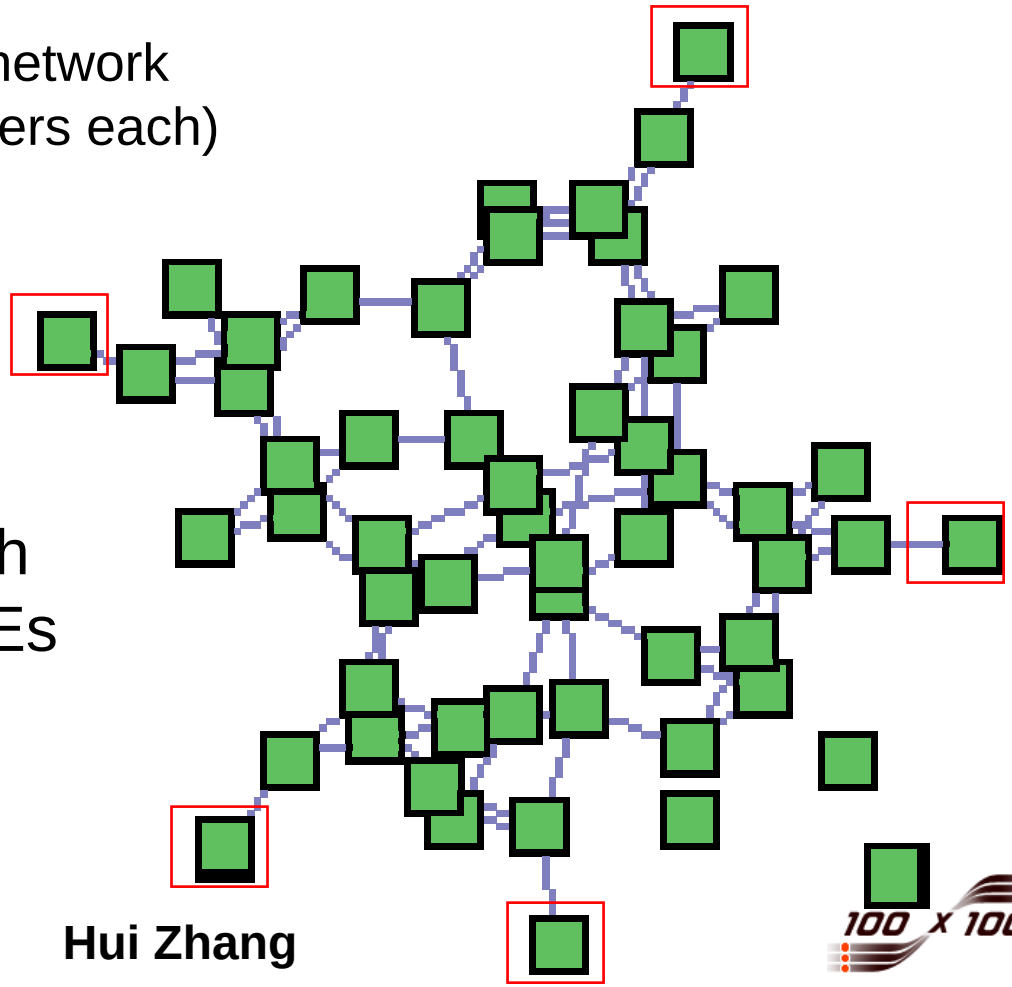
We designed and built a prototype of the 4D Architecture

- **4D Architecture permits many designs – prototype is a single, simple design point**
- **Decision plane**
  - Contains logic to simultaneously compute routes and enforce reachability matrix
  - Multiple Decision Elements per network, using simple election protocol to pick master
- **Dissemination plane**
  - Uses source routes to direct control messages
  - Extremely simple, but can route around failed data links

Evaluation of the 4D Prototype

- **Evaluated using Emulab (www.emulab.net)**
  - Linux PCs used as routers (650 – 800MHz)
  - Tested on 9 enterprise network topologies (10-100 routers each)

Example network with
49 switches and 5 DEs



Hui Zhang

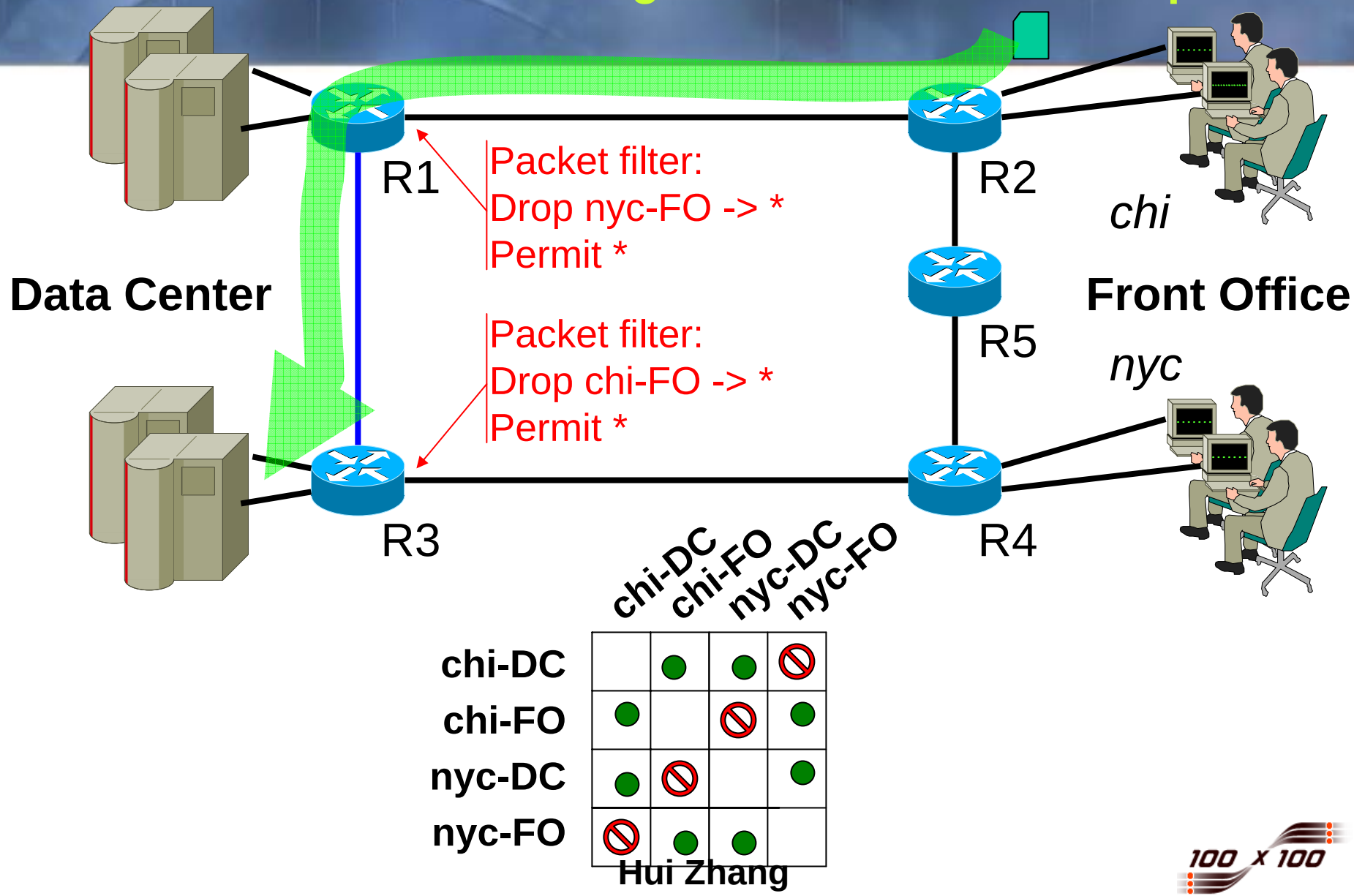
100 x 100

Performance of the 4D Prototype

Trivial prototype has performance comparable to well-tuned production networks

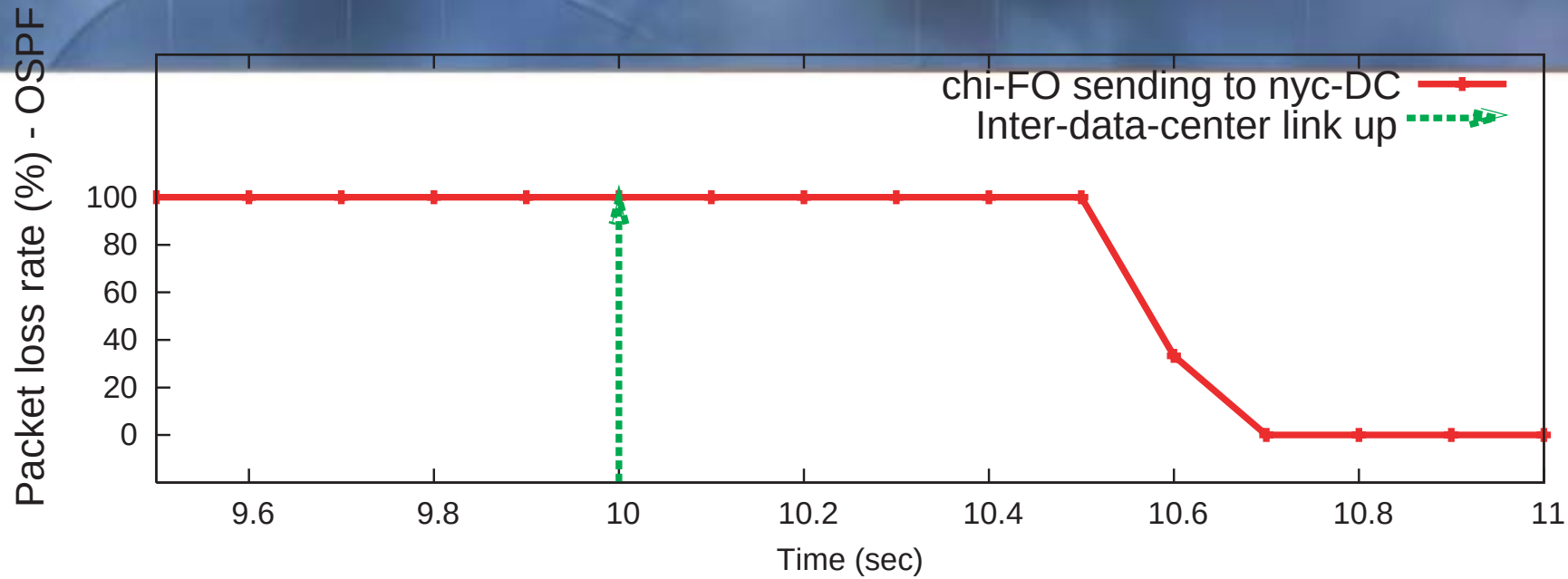
- **Recovers from single link failure in < 300 ms**
  - < 1 s response considered “excellent”
- **Survives failure of master Decision Element**
  - New DE takes control within 1 s
  - No disruption unless second fault occurs
- **Gracefully handles complete network partitions**
  - Less than 1.5 s of outage

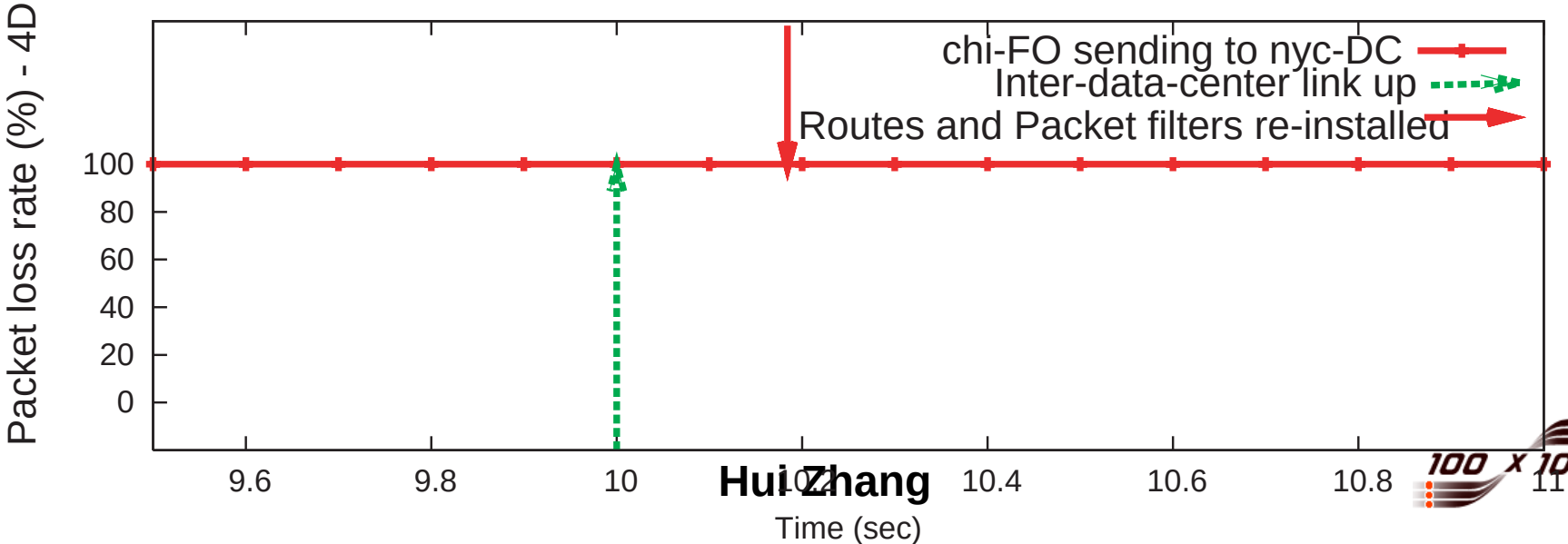
4D Makes Network Management & Control Error-proof



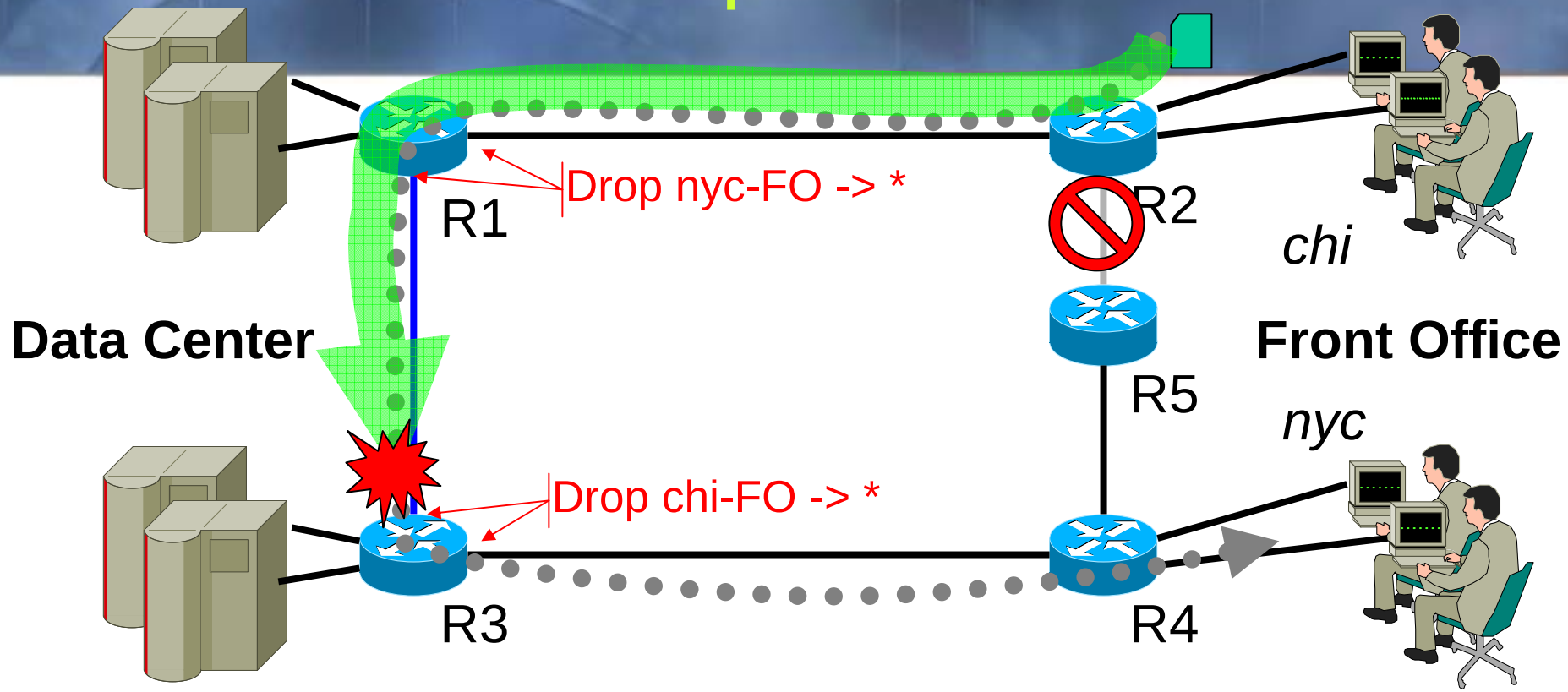
Hui Zhang

Prohibiting Packets from chi-FO to nyc-DC

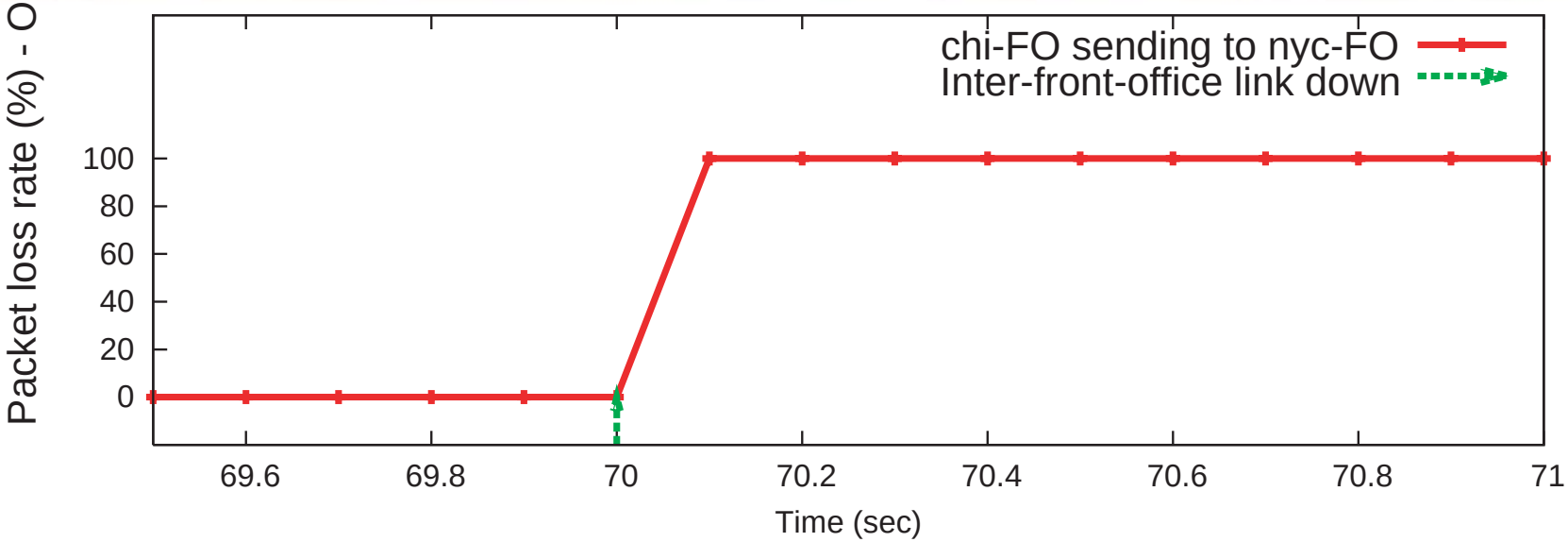


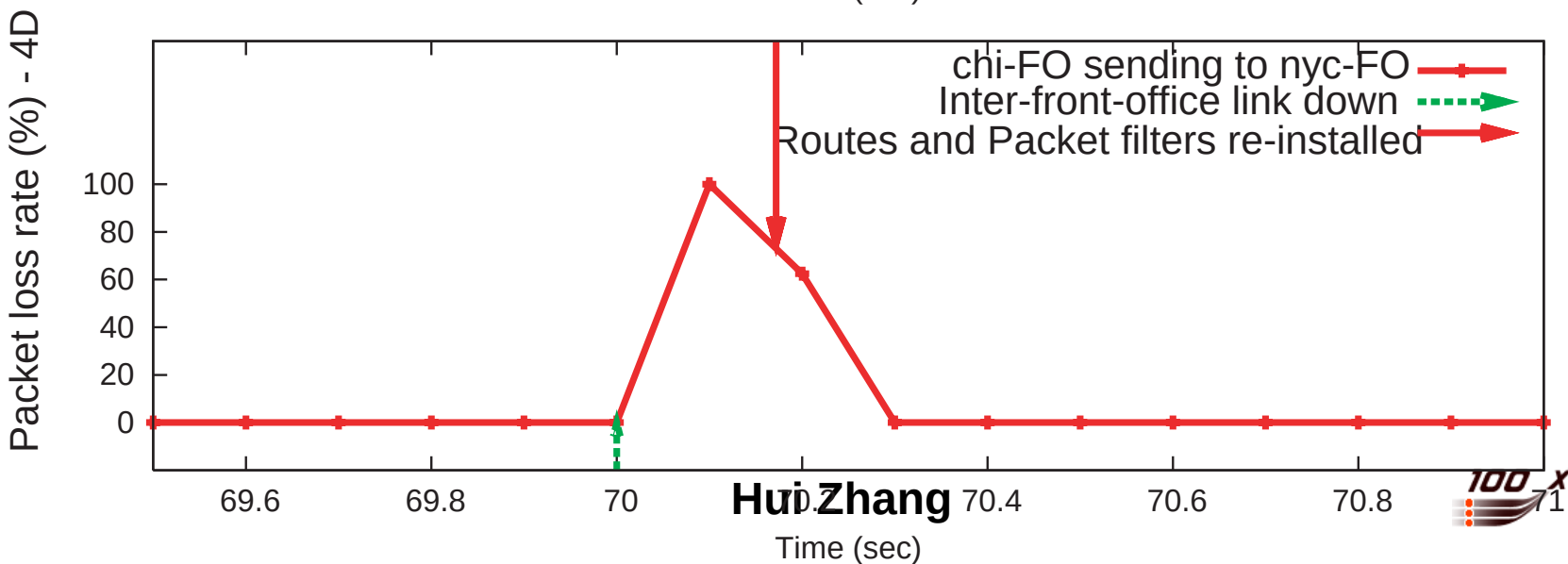


4D Makes Network Management & Control Error-proof



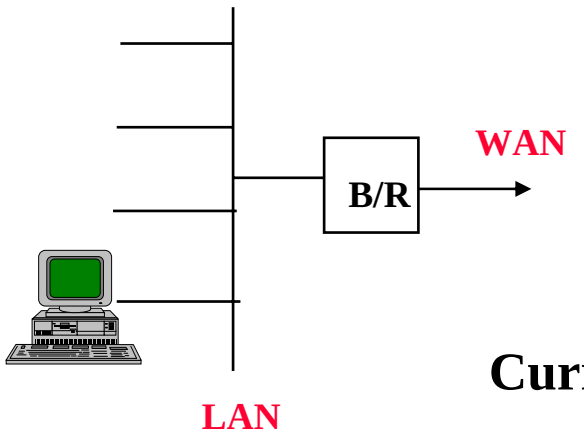
Allowing Packets from chi-FO to nyc-FO





Learning from Ethernet Evolution Experience

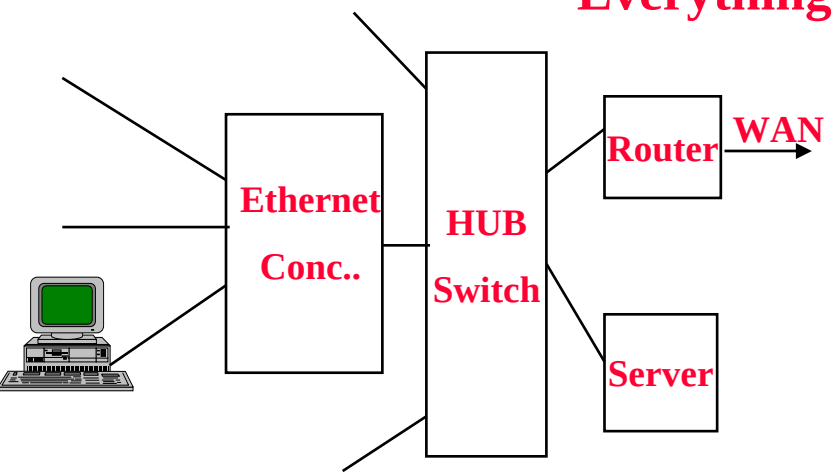
Early Implementations



- Bus-based Local Area Network
- Collision Domain, CSMA/CD
- Bridges and Repeaters for distance/capacity extension
- 1-10Mbps: coax, twisted pair (10BaseT)

Current Implementations:

Everything Changed Except Name and Framing



- Switched solution
- Little use for collision domains
- 80% of traffic leaves the LAN
- Servers, routers 10 x station speed
- 10/100/1000 Mbps, 10gig coming: Copper, Fiber

Hui Zhang

Control Plane: The Key Leverage Point

- **Great Potential: control plane determines the behavior of the network**
  - Reaction to events, reachability, services
- **Great Opportunities**
  - A radical clean-slate control plane can be deployed
    - Agnostic to packet format: IPv4/v6, ethernet
    - No changes to end-system software
  - Control plane is the nexus of network evolution
    - Changing the control plane logic can smooth transitions in network technologies and architectures

4D Supports Network Evolution & Expansion

- **Decision logic can be upgraded as needed**
  - No need for update of distributed protocols implemented in software distributed on every router
- **Decision elements can be upgraded as needed**
  - Network expansion requires changes only to DEs, not every router

Related Work

- **Separation of forwarding elements and control elements**
  - IETF: FORCES, GSMP, GMPLS
  - SoftRouter [Lakshman]
- **Driving network operation from network-wide views**
  - Traffic Engineering, Traffic Matrix computation
- **Centralization of decision making logic**
  - RCP [Feamster], PCE [Farrel]
  - SS7 [Ma Bell]

Summary

- Internet and IP have been a great success, and will continue to be more successful for years to come
- Never too late to think the next big thing
- Clean Slate Design could be a powerful research paradigm
- Control/management plane is where the problems and opportunities lie

Can We (Researchers) Make a Difference in the Future?

- **Monopoly positions in all technology areas**
  - Microsoft in OS
  - Cisco in router
  - Intel in processor
  - Oracle in Database
- **People are usually**
  - too optimistic in prediction of two years out, but
  - too pessimistic in prediction of five or ten years out

Characteristics of Big Bet Research

- **Visionary Ideas Carrying Intellectual Risk**
- **Can't Predict Outcomes in Advance**
  - The Christopher Columbus Effect

Randy Bryant: Dean of SCS, CMU
"Strategic Vision for CS in CMU"

Hui Zhang

100 x 100

Lead Dog Benefit

- **Other dogs see the same view**
  - the rear end of the dog ahead

Summary

- **Networks must meet many different types of objectives**
  - Security, traffic engineering, robustness
- **Today, objectives met using control plane mechanisms**
  - Results in complicated distributed system
  - Ripe with opportunities to set time-bombs
- **Refactoring into a 4D Architecture very promising**
  - Separates protocol issues from decision-making issues
  - Eliminates duplicate logic and simplifies network
  - Enables new capabilities, like joint control
  - Facilitate network evolution