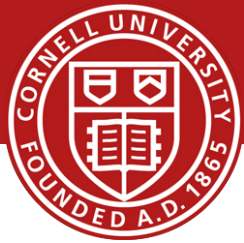


Data Center Traffic and Measurements

Hakim Weatherspoon

Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

November 10, 2014

Slides from SIGCOMM Internet Measurement Conference (IMC) 2010 presentation of "Analysis and Network Traffic Characteristics of Data Centers in the wild"

Goals for Today

- Analysis and Network Traffic Characteristics of Data Centers in the wild
  - T. Benson, A. Akella, and D. A. Maltz. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (IMC), pp. 267-280. ACM, 2010.

The Importance of Data Centers

- "***A 1-millisecond advantage in trading applications can be worth $100 million a year to a major brokerage firm***"

- Internal users
  - Line-of-Business apps
  - Production test beds
- External users
  - Web portals
  - Web services
  - Multimedia applications
  - Chat/IM

The Case for Understanding Data Center Traffic

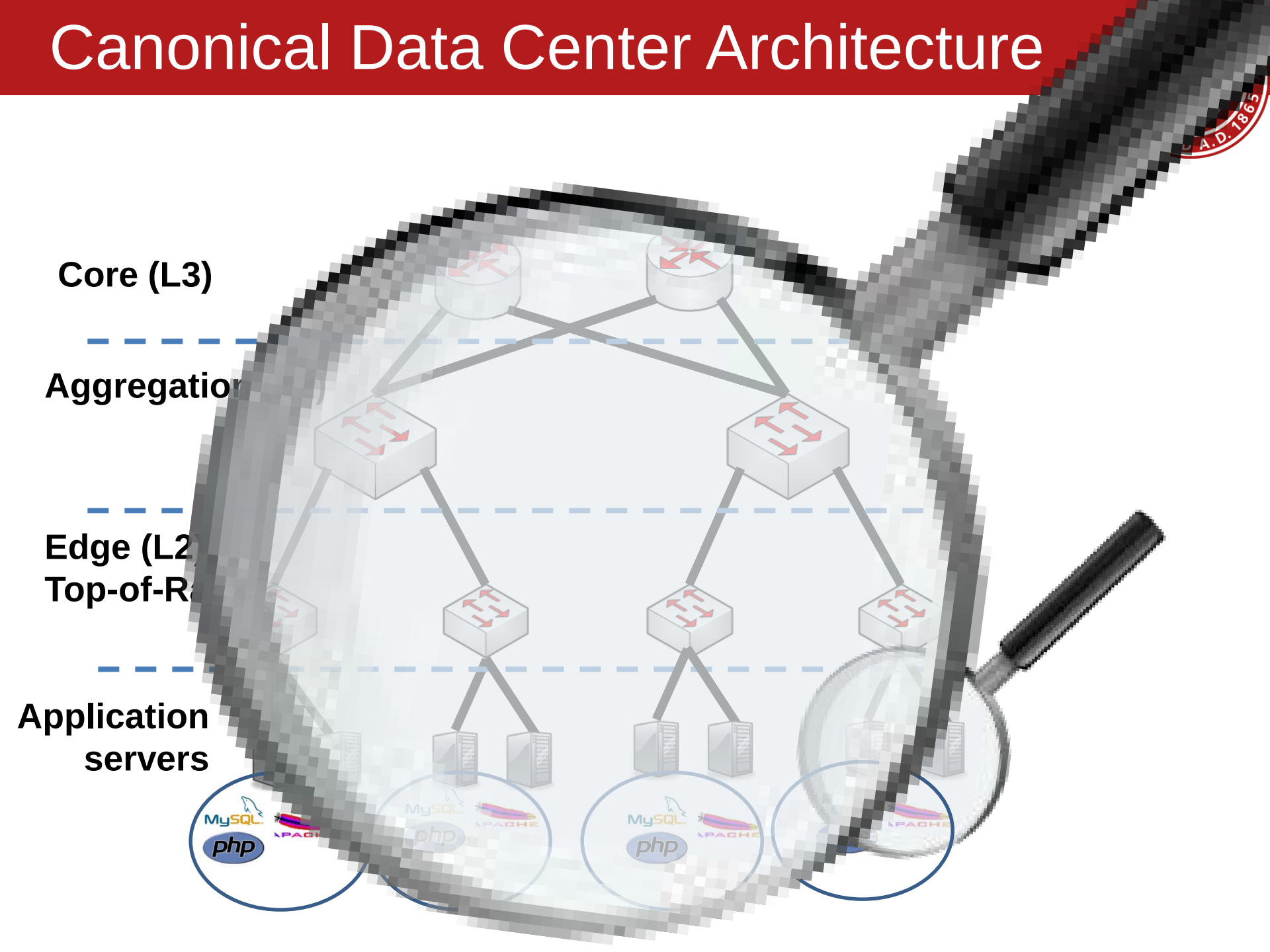
- Better understanding → better techniques
- Better traffic engineering techniques
  - Avoid data losses
  - Improve app performance
- Better Quality of Service techniques
  - Better control over jitter
  - Allow multimedia apps
- Better energy saving techniques
  - Reduce data center's energy footprint
  - Reduce operating expenditures

- Initial stab→ network level traffic + app relationships

Take aways and Insights Gained

- 75% of traffic stays within a rack (Clouds)
  - Applications are not uniformly placed
- Half packets are small (< 200B)
  - Keep alive integral in application design
- At most 25% of core links highly utilized
  - Effective routing algorithm to reduce utilization
  - Load balance across paths and migrate VMs
- Questioned popular assumptions
  - Do we need more bisection? No
  - Is centralization feasible? Yes

Canonical Data Center Architecture



Dataset: Data Centers Studied

- 10 data centers
- 3 classes
  - Universities
  - Private enterprise
  - Clouds
- Internal users
  - Univ/priv
  - Small
  - Local to campus
- External users
  - Clouds
  - Large
  - Globally diverse

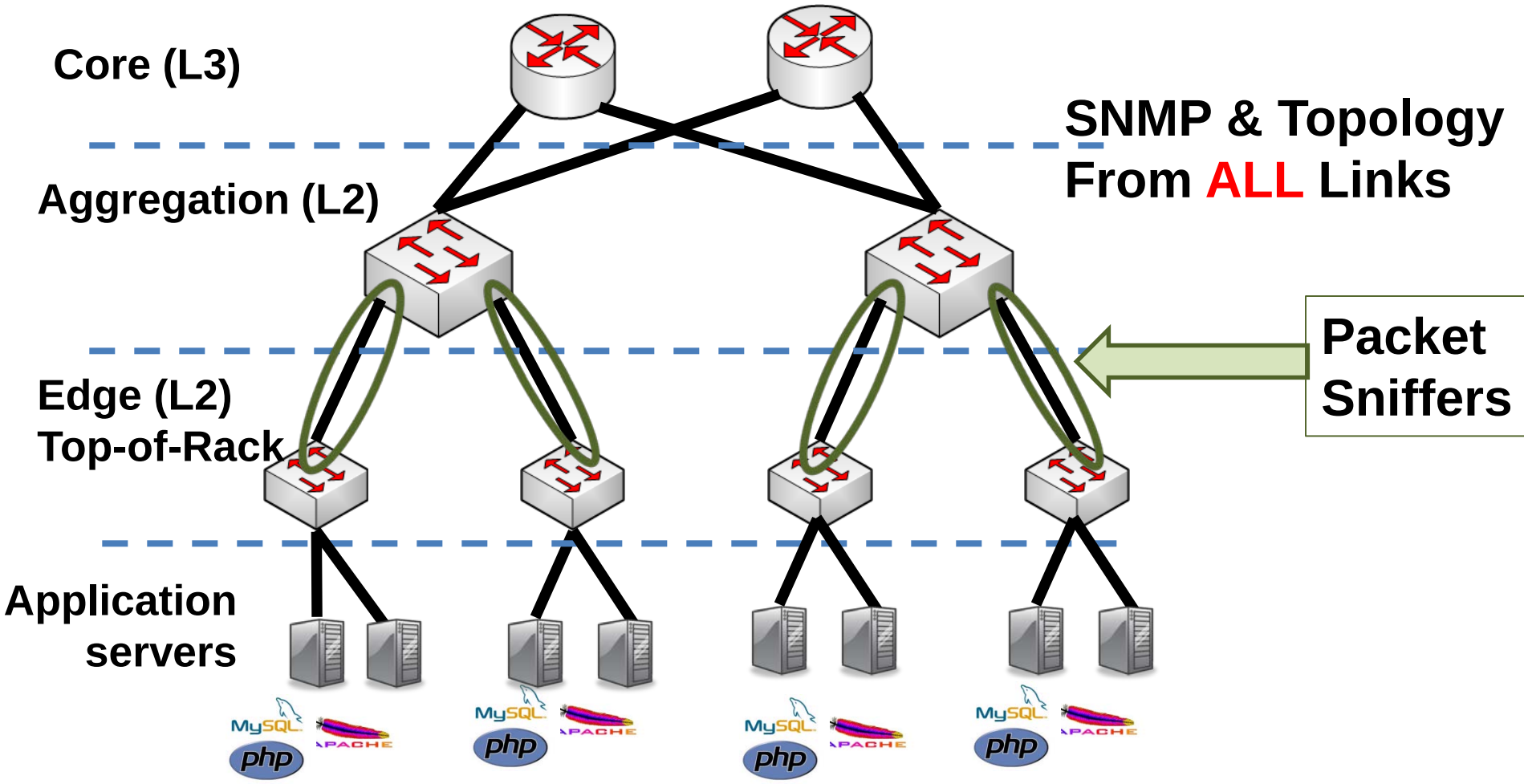
DC Role	DC Name	Location	Number Devices
Universities	EDU1	US-Mid	22
	EDU2	US-Mid	36
	EDU3	US-Mid	11
Private Enterprise	PRV1	US-Mid	97
	PRV2	US-West	100
Commercial Clouds	CLD1	US-West	562
	CLD2	US-West	763
	CLD3	US-East	612
	CLD4	S. America	427
	CLD5	S. America	427

Dataset: Collection

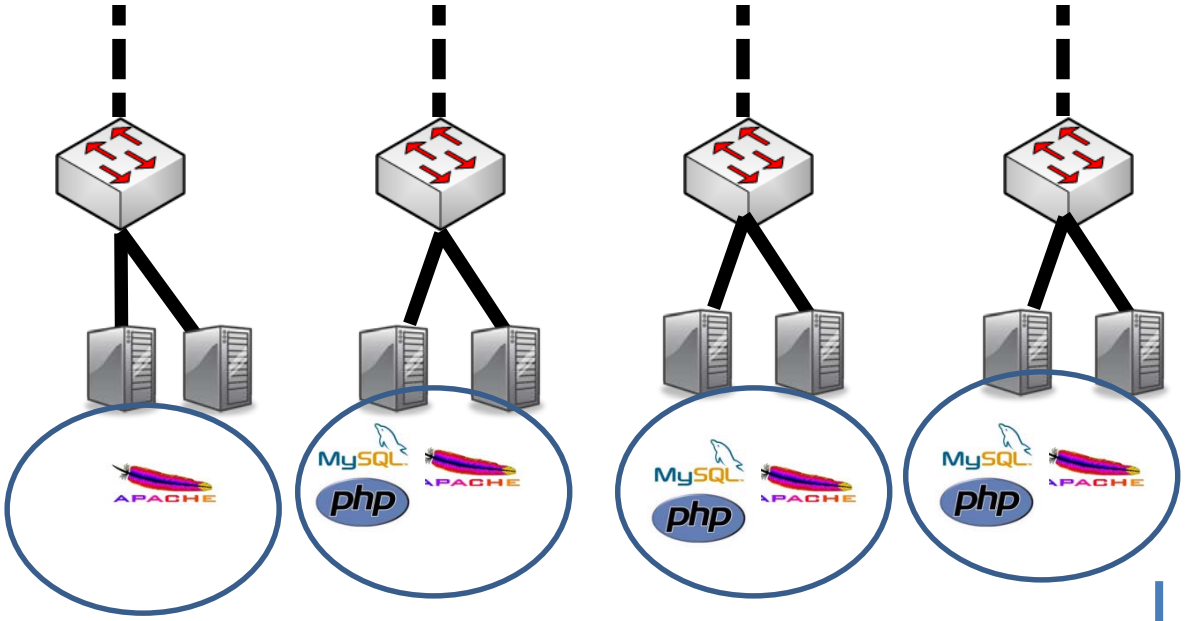
- SNMP
  - Poll SNMP MIBs
  - Bytes-in/bytes-out/discards
  - > 10 Days
  - Averaged over 5 mins
- Packet Traces
  - Cisco port span
  - 12 hours
- Topology
  - Cisco Discovery Protocol

DC Name	SNMP	Packet Traces	Topology
EDU1	Yes	Yes	Yes
EDU2	Yes	Yes	Yes
EDU3	Yes	Yes	Yes
PRV1	Yes	Yes	Yes
PRV2	Yes	Yes	Yes
CLD1	Yes	No	No
CLD2	Yes	No	No
CLD3	Yes	No	No
CLD4	Yes	No	No
CLD5	Yes	No	No

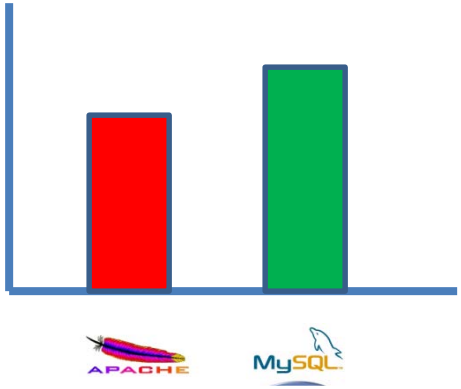
Canonical Data Center Architecture



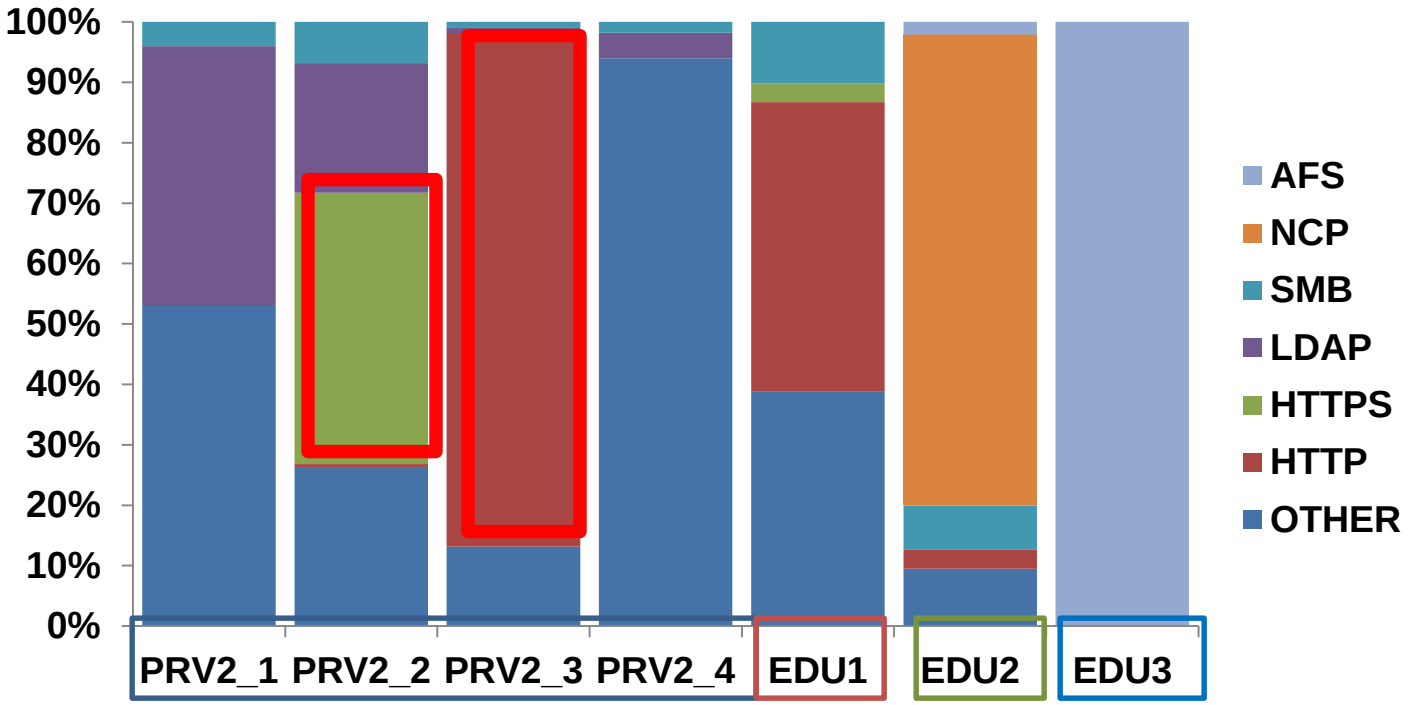
Applications



- Start at bottom
  - Analyze running applications
  - Use packet traces
- BroID tool for identification
  - Quantify amount of traffic from each app



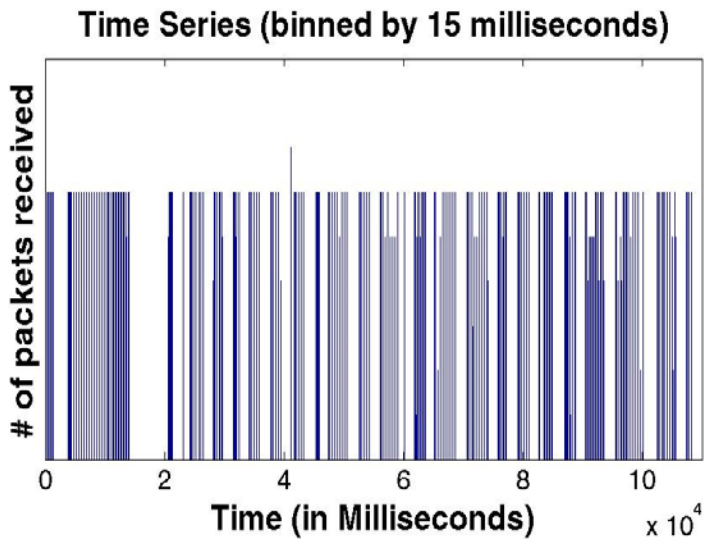
Applications

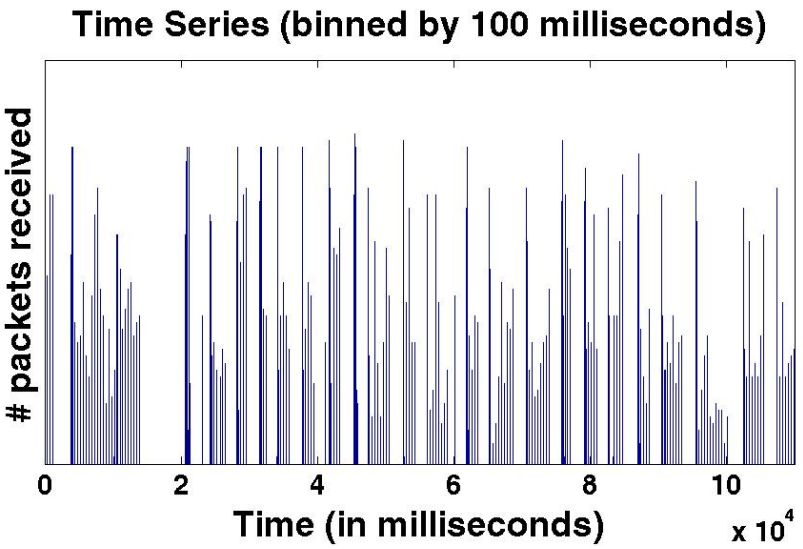


- Differences between various bars
- Clustering of applications
  - PRV2_2 hosts secured portions of applications
  - PRV2_3 hosts unsecure portions of applications

Analyzing Packet Traces

- Transmission patterns of the applications
- Properties of packet crucial for
  - Understanding effectiveness of techniques

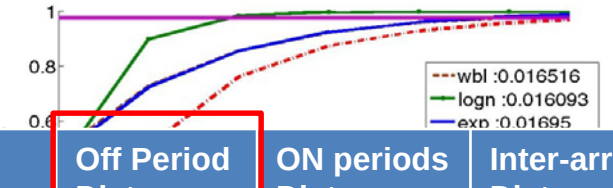




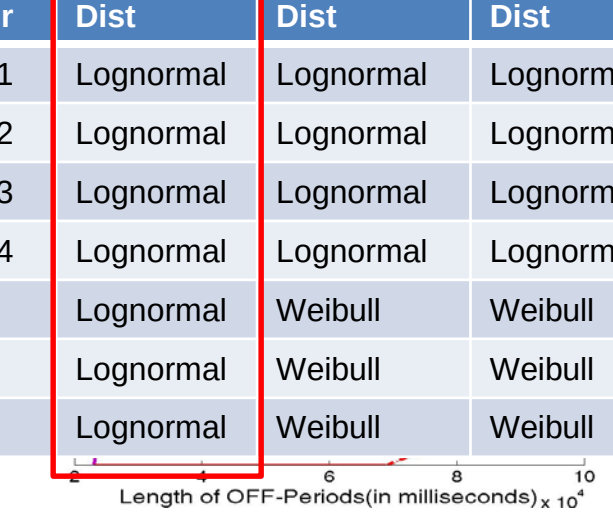
- ON-OFF traffic at edges
  - Binned in 15 and 100 m. secs
  - We observe that ON-OFF persists

Data Center Traffic is Bursty

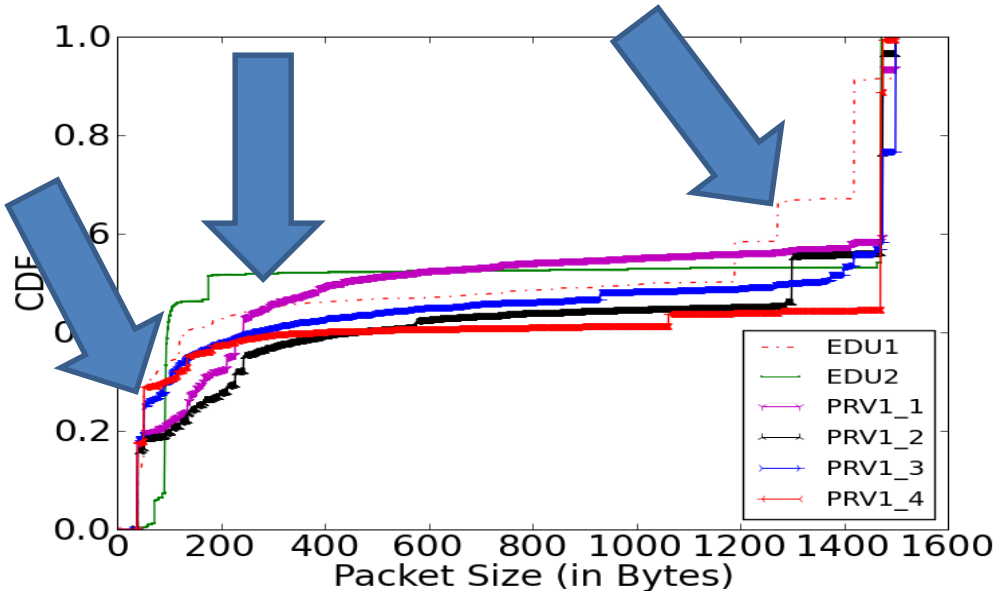
- Understanding arrival process
  - Range of acceptable models
- What is the arrival process?
  - **Heavy-tail** for the 3 distributions
    - ON, OFF times, Inter-arrival,
  - **Lognormal** across all data centers
- Different from Pareto of WAN
  - Need new models



Data Center	Off Period Dist	ON periods Dist	Inter-arrival Dist
Prv2_1	Lognormal	Lognormal	Lognormal
Prv2_2	Lognormal	Lognormal	Lognormal
Prv2_3	Lognormal	Lognormal	Lognormal
Prv2_4	Lognormal	Lognormal	Lognormal
EDU1	Lognormal	Weibull	Weibull
EDU2	Lognormal	Weibull	Weibull
EDU3	Lognormal	Weibull	Weibull

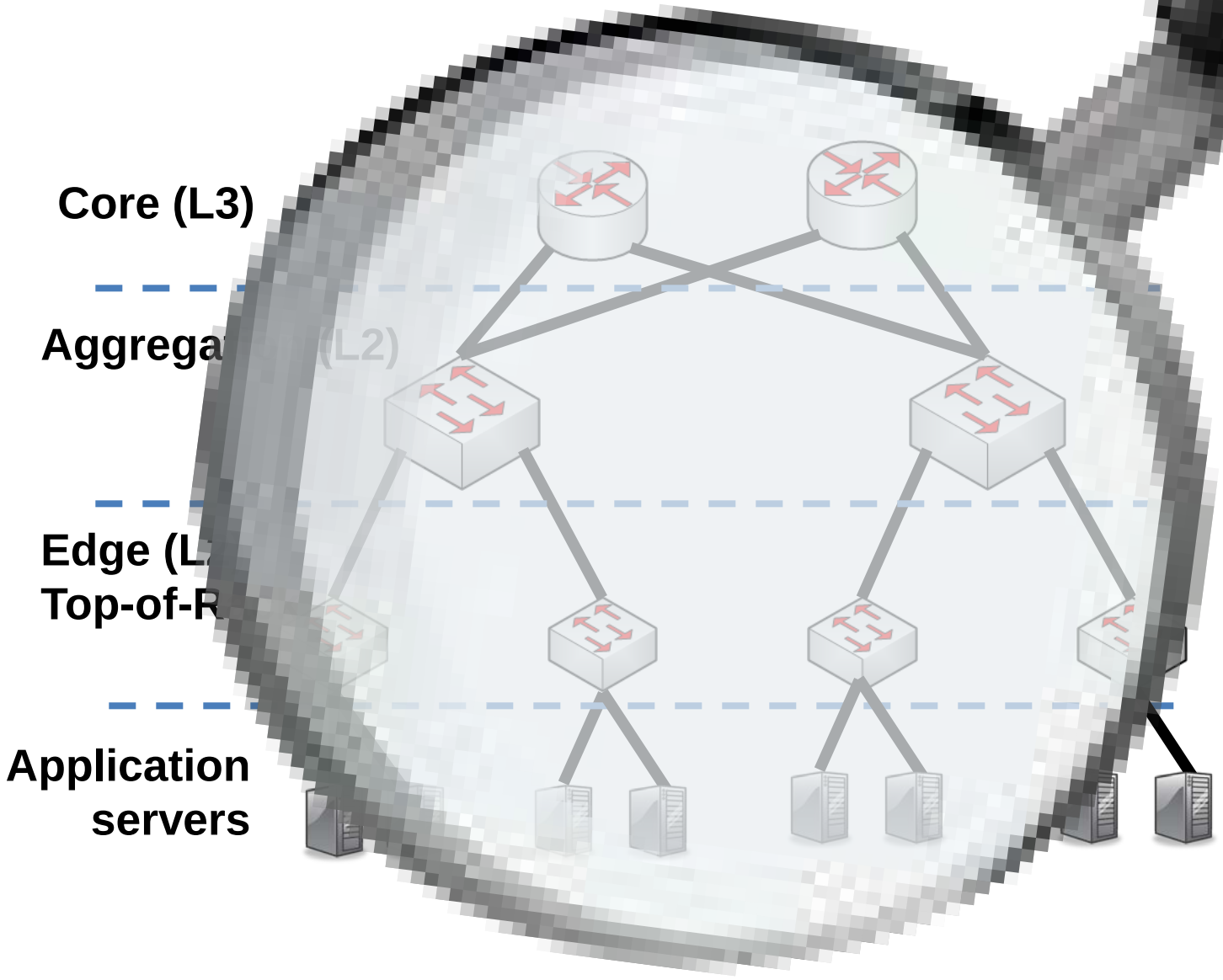


Packet Size Distribution



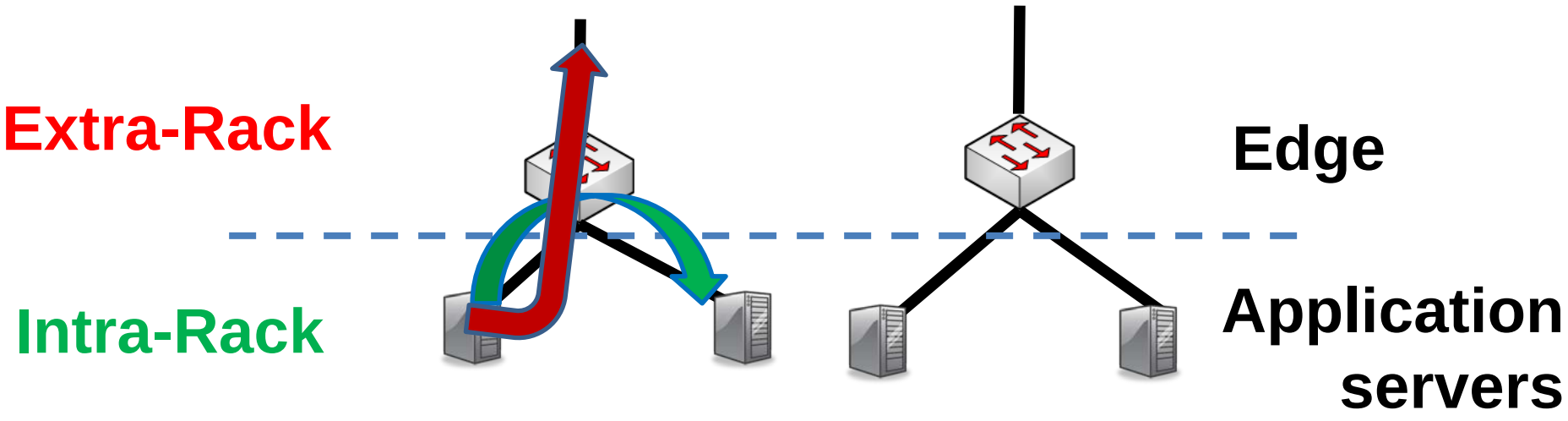
- Bimodal (200B and 1400B)
- Small packets
  - TCP acknowledgements
  - Keep alive packets
- Persistent connections → important to apps

Canonical Data Center Architectur



Intra-Rack Versus Extra-Rack

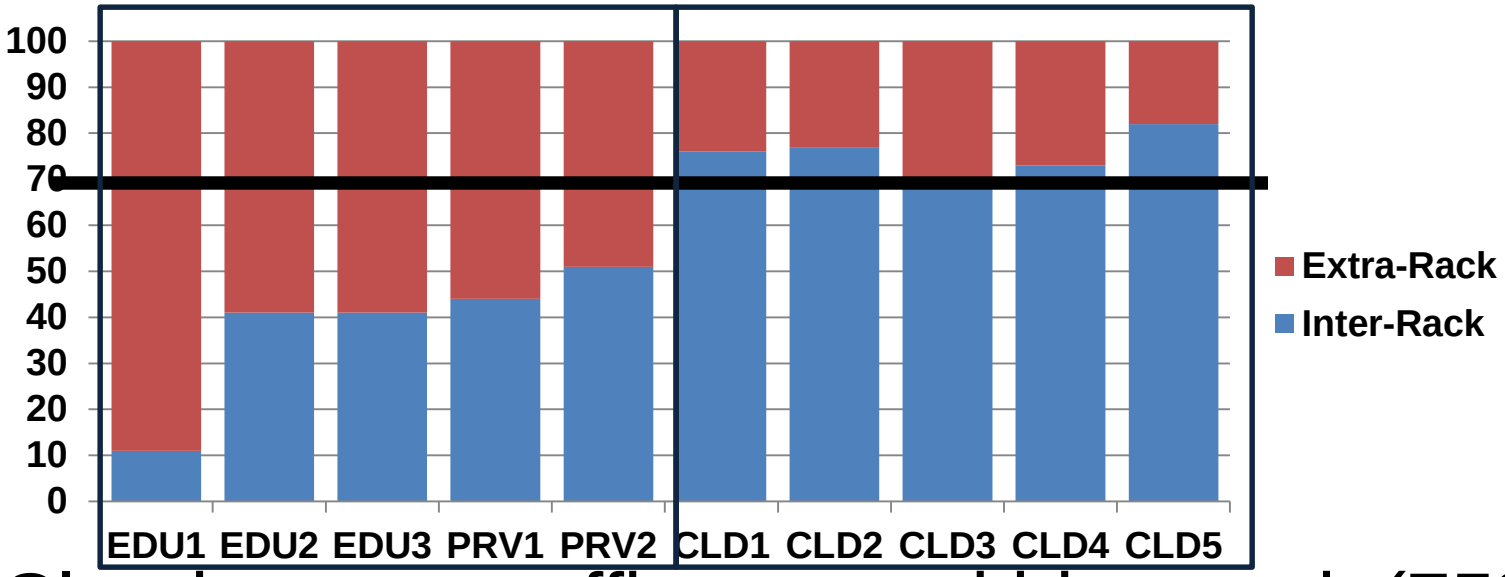
- Quantify amount of traffic using interconnect
  - Perspective for interconnect analysis



Extra-Rack = Sum of Uplinks

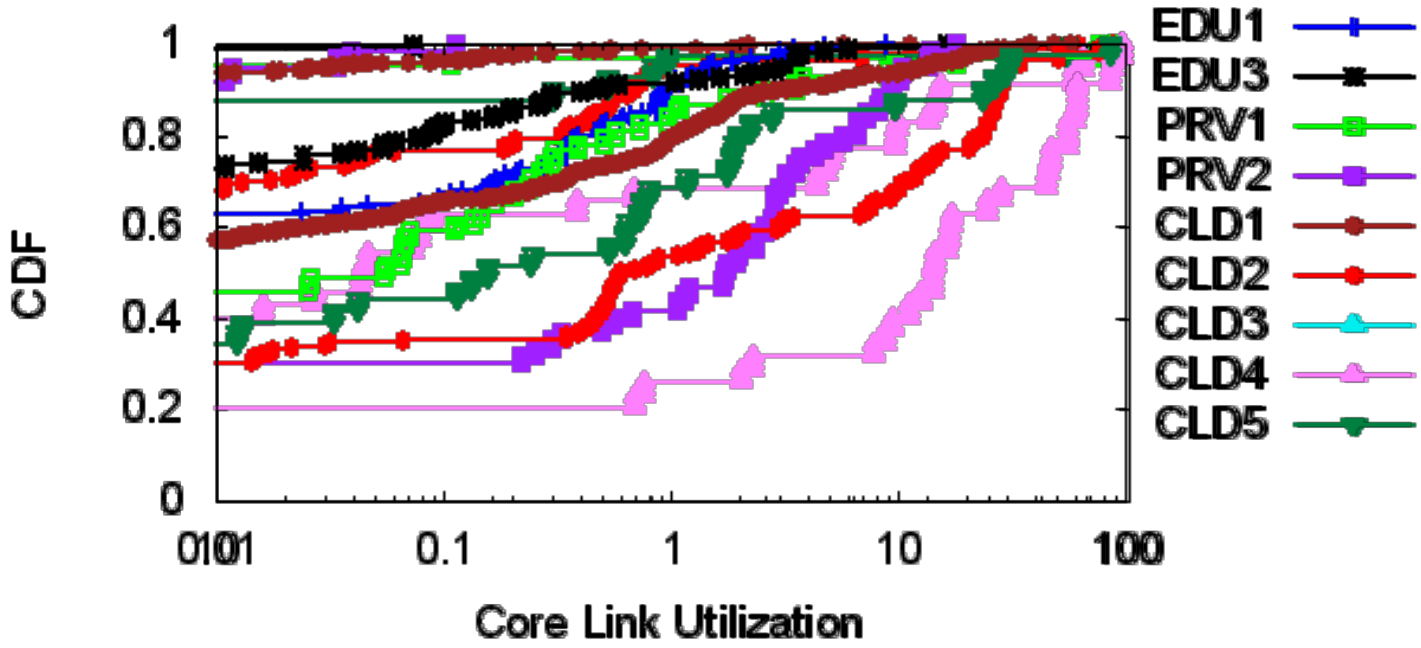
Intra-Rack = Sum of Server Links – Extra-Rack

Intra-Rack Versus Extra-Rack Results



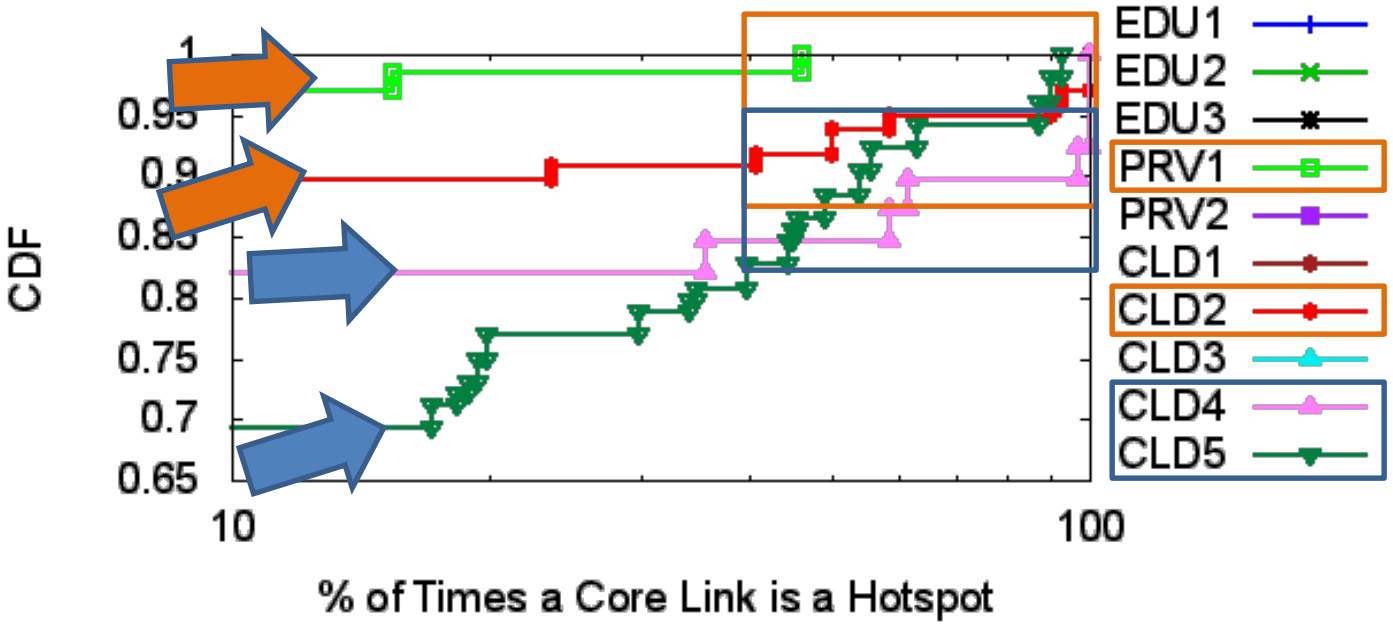
- Clouds: most traffic stays within a rack (75%)
  - Colocation of apps and dependent components
- Other DCs: > 50% leaves the rack
  - Un-optimized placement

Extra-Rack Traffic on DC Interconnect



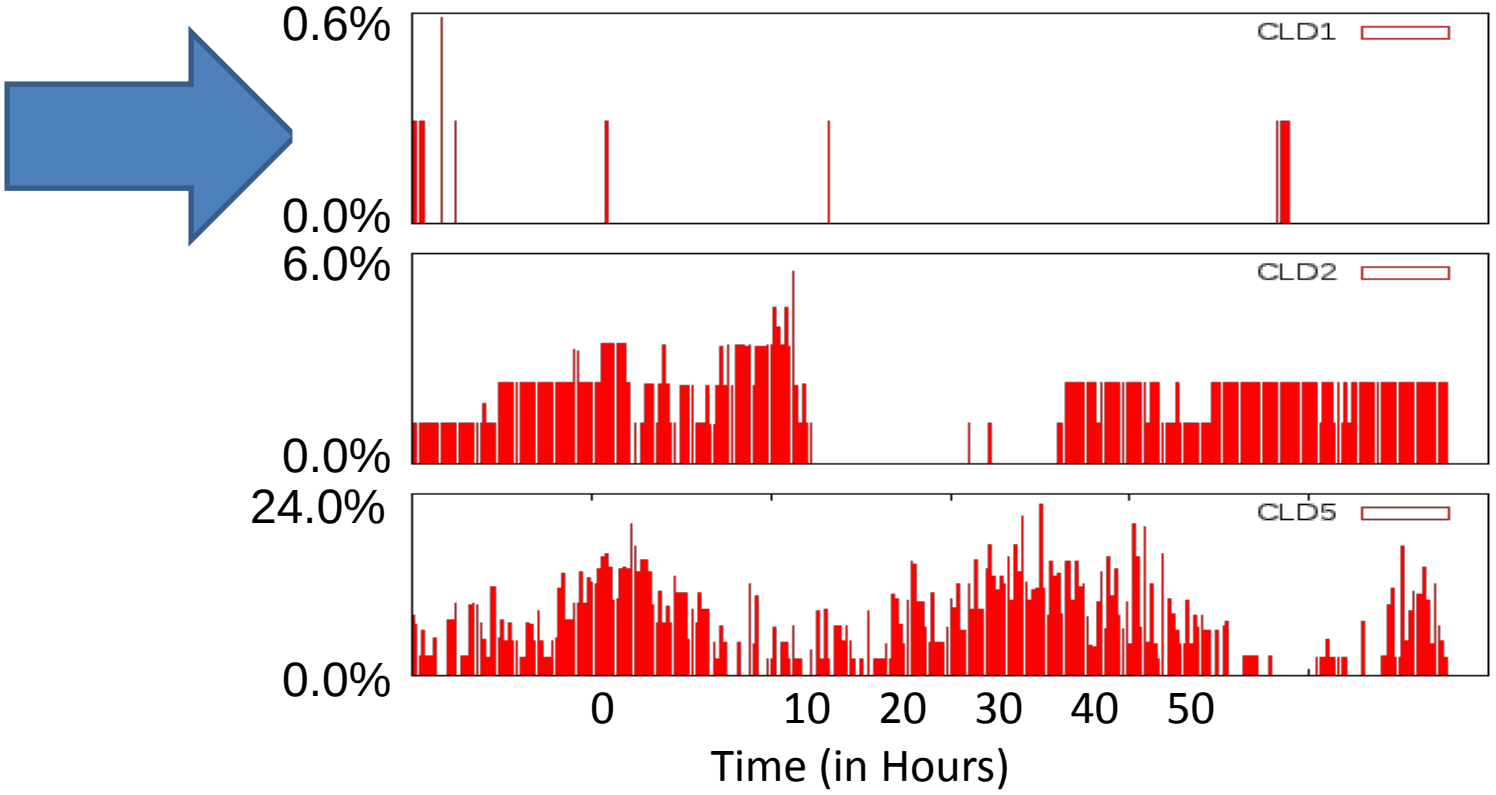
- Utilization: core > agg > edge
  - Aggregation of many unto few
- Tail of core utilization differs
  - Hot-spots → links with > 70% util
  - Prevalence of hot-spots differs across data centers

Persistence of Core Hot-Spots



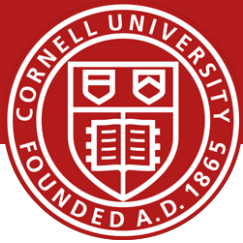
- Low persistence: PRV2, EDU1, EDU2, EDU3, CLD1, CLD3
- High persistence/low prevalence: PRV1, CLD2
  - 2-8% are hotspots > 50%
- High persistence/high prevalence: CLD4, CLD5
  - 15% are hotspots > 50%

Prevalence of Core Hot-Spots



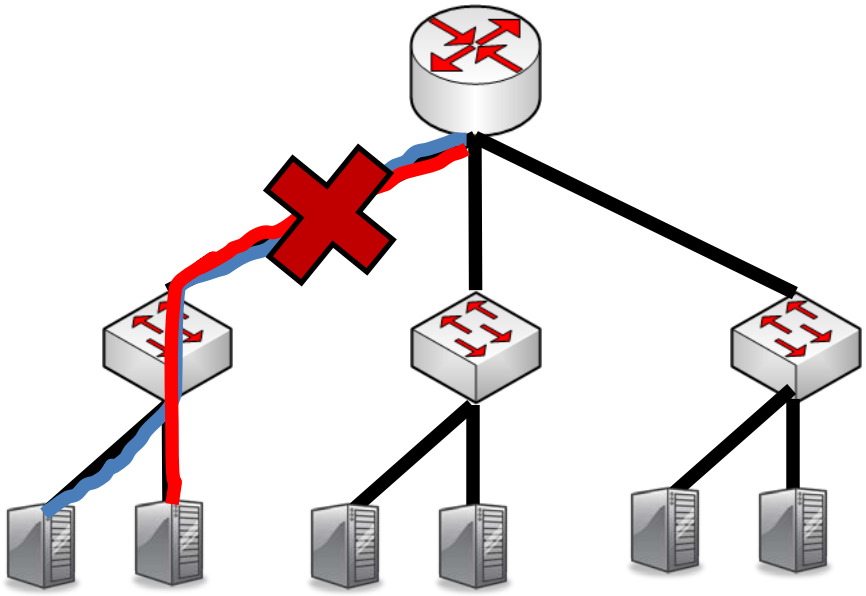
- Low persistence: very few concurrent hotspots
- High persistence: few concurrent hotspots
- High prevalence: < 25% are hotspots at any time

Observations from Interconnect



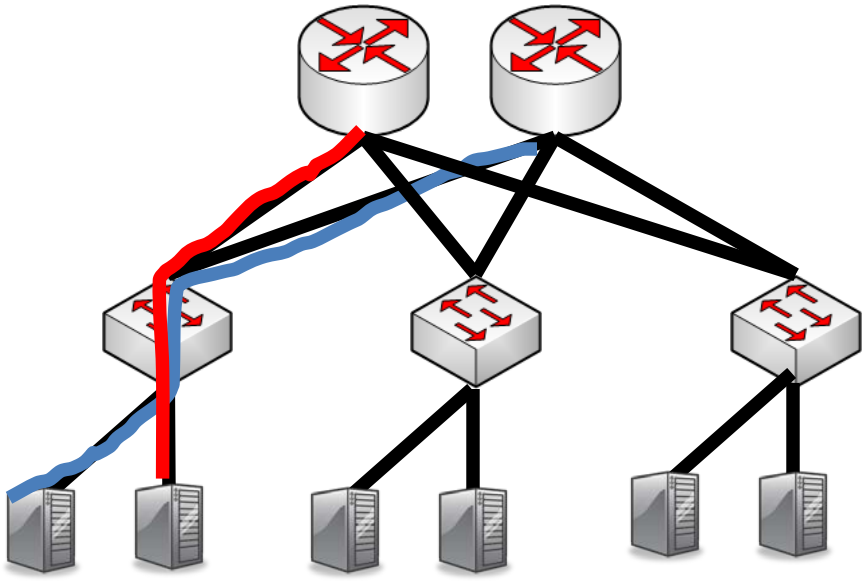
- Links utils low at edge and agg
- Core most utilized
  - Hot-spots exists (> 70% utilization)
  - < 25% links are hotspots
  - Loss occurs on less utilized links (< 70%)
    - Implicating momentary bursts
- Time-of-Day variations exists
  - Variation an order of magnitude larger at core
- Apply these results to evaluate DC design requirements

Assumption 1: Larger Bisection



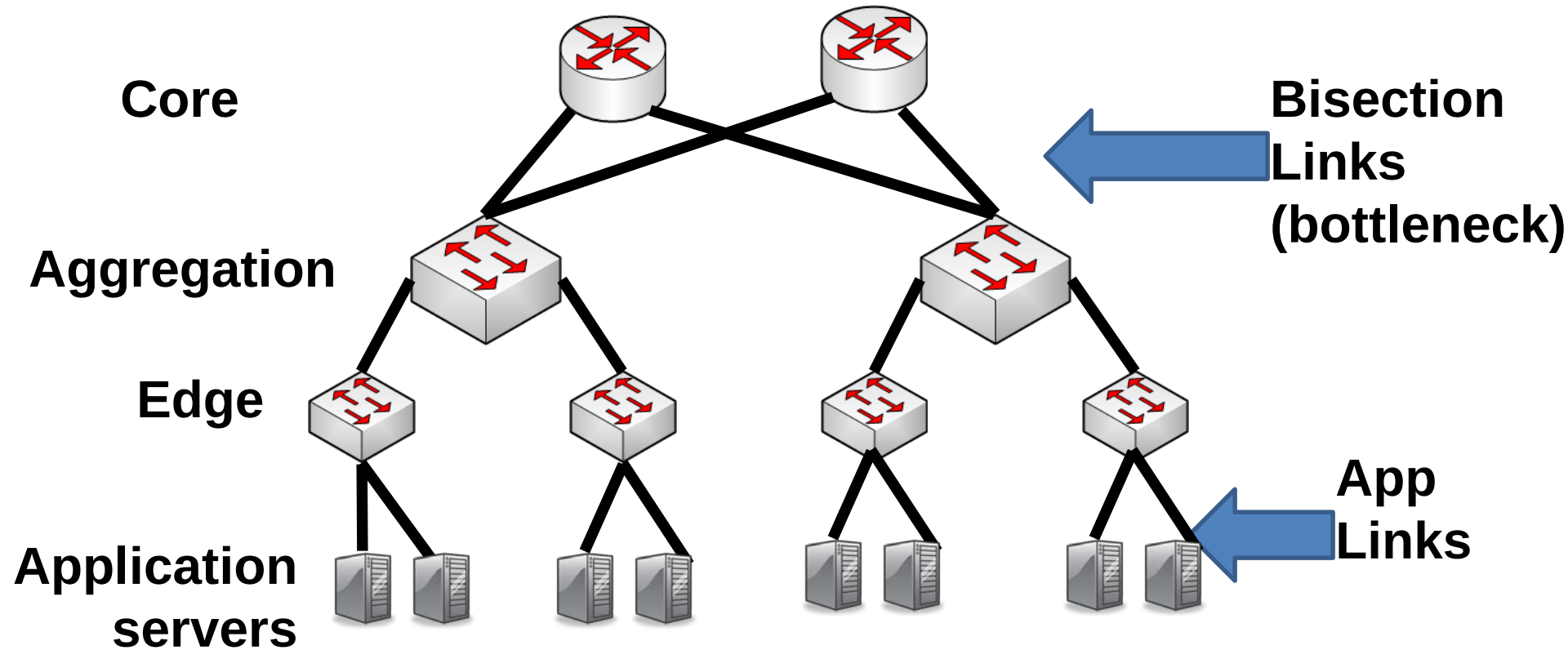
- Need for larger bisection
  - VL2 [Sigcomm ‘09], Monsoon [Presto ‘08],Fat-Tree [Sigcomm ‘08], Portland [Sigcomm ‘09], Hedera [NSDI ’10]
  - Congestion at oversubscribed core links

Argument for Larger Bisection



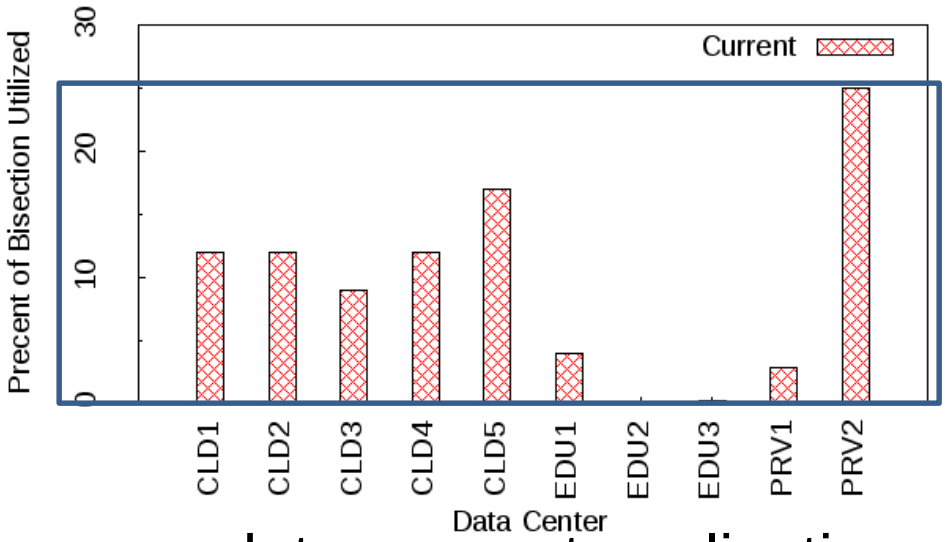
- Need for larger bisection
  - VL2 [Sigcomm ‘09], Monsoon [Presto ‘08],Fat-Tree [Sigcomm ‘08], Portland [Sigcomm ‘09], Hedera [NSDI ’10]
  - Congestion at oversubscribed core links
  - Increase core links and eliminate congestion

Calculating Bisection Bandwidth



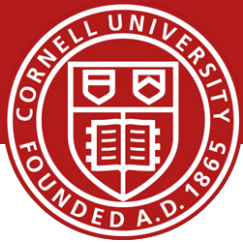
If $\left(\frac{\Sigma \text{ traffic (App)}}{\Sigma \text{ capacity(Bisection}} \right) > 1$ then more device are needed at the bisection

Bisection Demand



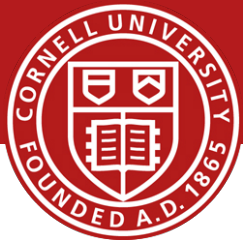
- Given our data: current applications and DC design
  - NO, more bisection is not required
  - Aggregate bisection is only 30% utilized
- Need to better utilize existing network
  - Load balance across paths
  - Migrate VMs across racks

Insights Gained



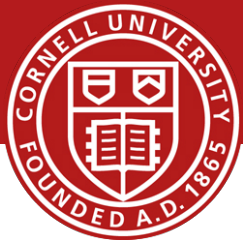
- 75% of traffic stays within a rack (Clouds)
  - Applications are not uniformly placed
- Half packets are small (< 200B)
  - Keep alive integral in application design
- At most 25% of core links highly utilized
  - Effective routing algorithm to reduce utilization
  - Load balance across paths and migrate VMs
- Questioned popular assumptions
  - Do we need more bisection? No
  - Is centralization feasible? Yes

Related Works



- IMC ‘09 [Kandula`09]
  - Traffic is unpredictable
  - Most traffic stays within a rack
- Cloud measurements [Wang’10,Li’10]
  - Study application performance
  - End-2-End measurements

Before Next time



- Project Interim report
  - **Due Monday, November 24.**
  - And meet with groups, TA, and professor
- Fractus Upgrade: Should be back online

- ***Required review and reading for Wednesday, November 12***
  - SoNIC: Precise Realtime Software Access and Control of Wired Networks, K. Lee, H. Wang and H. Weatherspoon. USENIX symposium on Networked Systems Design and Implementation (NSDI), April 2013, pages 213-225.
  - https://www.usenix.org/system/files/conference/nsdi13/nsdi13-final138.pdf

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule