# Datacenter Traffic Measurement and Classification

**Speaker: Lin Wang**

Research Advisor: Biswanath Mukherjee

UCDAVIS

# Datacenter Traffic Measurement and Analysis

- **Data Collection**
  - Collect network events from each of 1500 servers
  - For over two months.
- **Traffic Characteristics**
  - Server pairs within the same rack more likely to exchange more bytes.
  - 21% probability to exchange data within the same rack.
  - 0.5% probability to exchange data in different racks.

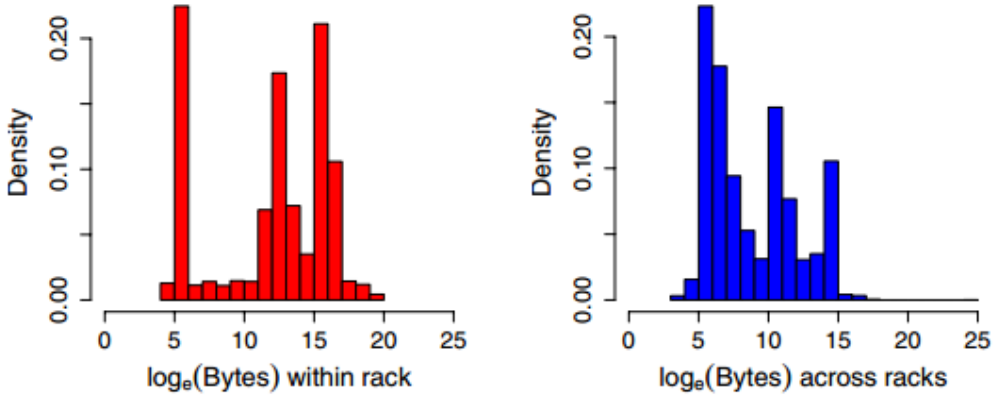



**Figure 3: How much traffic is exchanged between server**



Kandula S, Sengupta S, Greenberg A, et al. The nature of data center traffic: measurements & analysis[C]//Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, 2009: 202-208.

# Datacenter Traffic Measurement and Analysis

- **Traffic Characteristics**
  - Server either talks to almost all the other servers within the rack (the bump near 1 in figure left) or fewer than 25% of servers within the rack.
  - Server either doesn't talk to servers outside its rack (the spike at zero in figure right) or it talks to about 1-10% of outside servers.

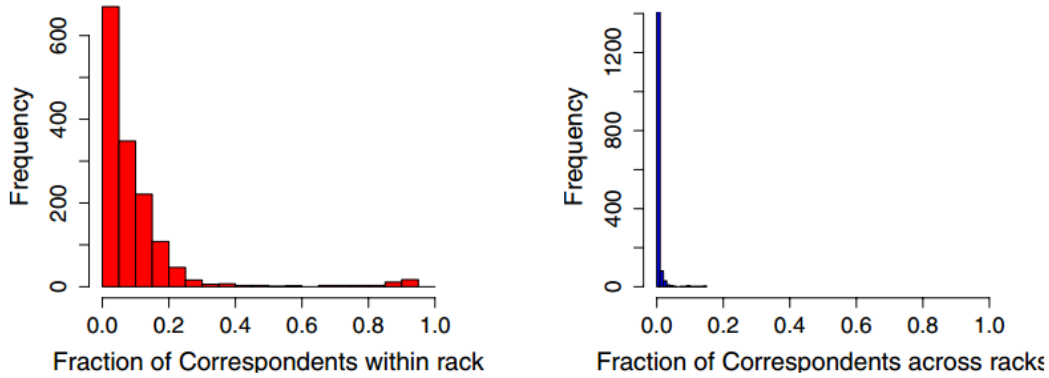



Figure 4: How many other servers does a server correspond

# Datacenter Traffic Measurement and Analysis

- **Traffic Characteristics**
  - Compare the rates of flows that overlap high utilization periods.
  - Rates do not change appreciably (see cdf below).
  - Errors(e.g. flow timeouts or failure) is not visible in flow rates.
  - Hence we correlate high utilization epochs directly with application level logs.

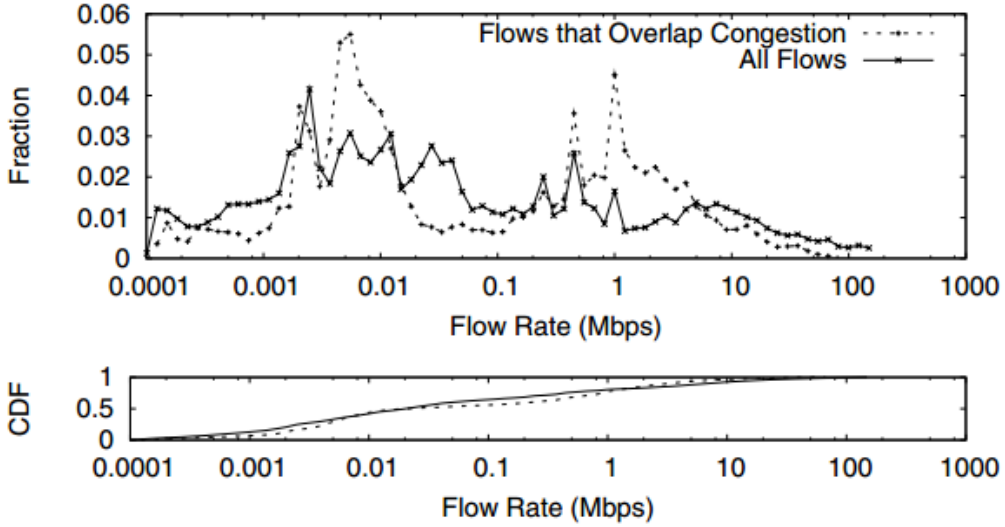



**Figure 7: Comparing rates of flows that overlap congestion with rates of all flows.**

# Datacenter Traffic Measurement and Analysis

- **Traffic Characteristics**
  - Traffic mix changes frequently.
  - The figure plots the durations of million flows (a day's worth of flows) in the cluster.
  - Most flows come and go (80% last less than 10s) and there are few long running flows (less than 0.1% last longer than 200s).

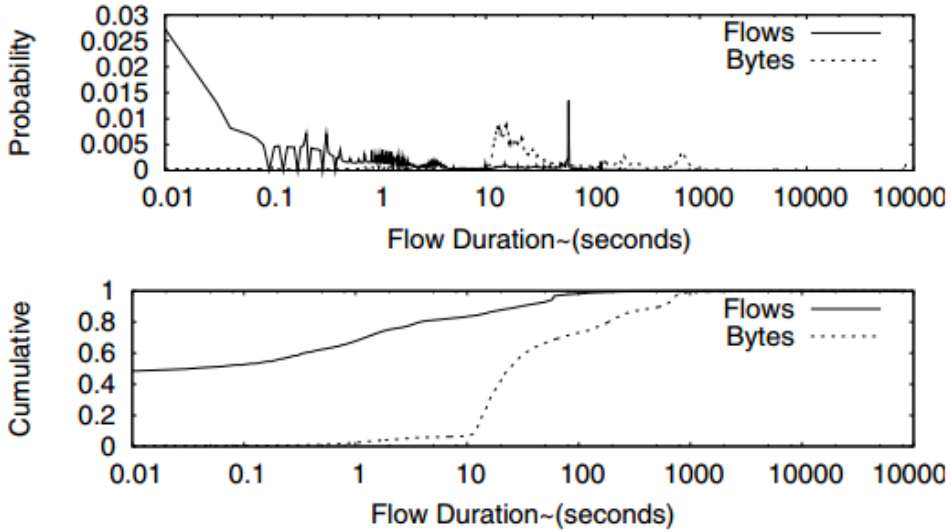


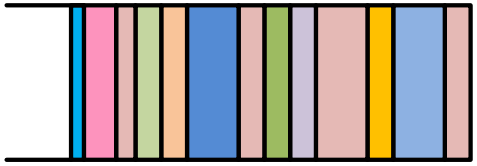

**Figure 9: More than 80% of the flows last less than ten seconds, fewer than .1% last longer than 200s and more than 50% of the bytes are in flows lasting less than 25s.**

UC DAVIS
UNIVERSITY OF CALIFORNIA

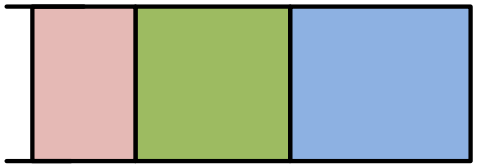
# Datacenter Traffic Classification

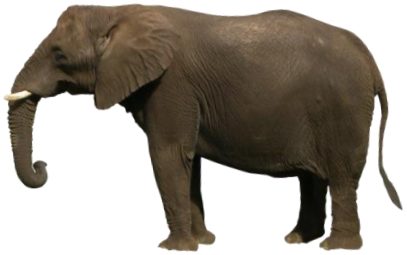
- **Why do we need to identify elephant flow?**
  - Previous paper shows that a large fraction of datacenter traffic is carried in a small fraction of flows.
  - 90% of the flows carry less than 1MB of data
  - >90% of bytes transferred are in flows greater than 100MB.
  - Hash-based flow forwarding techniques (e.g. Equal-Cost Multi-Path (ECMP) routing) works well only for mice flows and no elephant flows.

# Mice flow VS elephant flow



- Small size packet
- Short flow
- Large number
- Short-lived

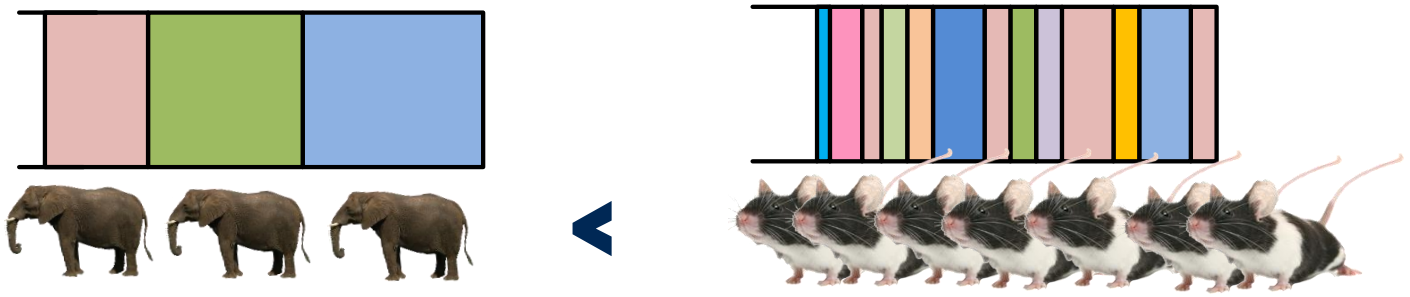




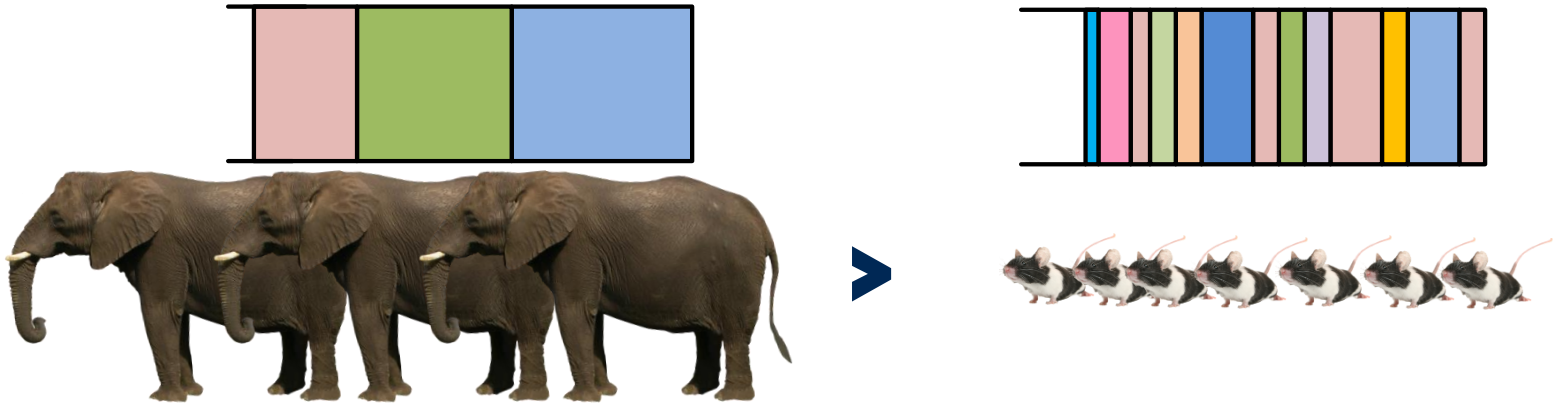
- Large size packet
- Large volume flow
- Small number
- Long lasting

# Mice flow VS elephant flow

- If we only care about the number of packets in the queue, elephant flow transmission is easy to be degraded.



- If we only care about the total size of packets in the queue, mice flow transmission is easy to be degraded.



# Datacenter Traffic Classification

- **Solution:**
  - Mahout, a low-overhead yet effective traffic management system.
  - End-host-based elephant detection.

- **Advantages of detecting elephant flow at end host.**
  - Network behavior of a flow is affected by how rapidly the end-point applications are generating data, and this is not biased by congestion in the network.
  - In contrast to in-network monitors, the end host OS has better visibility into the behavior of applications.
  - In datacenters, it is possible to augment the end host OS; this is enabled by the single administrative domain and software uniformity typical of modern datacenters.
  - Use very little overhead. In contrast, using an in-network mechanism to monitor is infeasible, even on an edge switch, and even more so on a core switch.



Curtis A R, Kim W, Yalagandula P. Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection[C]//INFOCOM, 2011 Proceedings IEEE. IEEE, 2011: 1629-1637.

# Datacenter Traffic Classification

- **Mahout algorithm:**
  - Usea shim layer in the end hosts to monitor the socket buffers.

**Algorithm 1** Pseudocode for end host shim layer

```
1: When sending a packet
2: if number of bytes in buffer ≥ threshold_elephant then
3:    / * Elephant flow */
4:    if last-tagged-time - now() ≥ T_tagperiod then
5:       set DS = 00001100
6:       last-tagged-time = now()
7:    end if
8: end if
```

# Datacenter Traffic Classification

- **Simulation parameters:**

| Parameter | Description | Value |
|---|---|---|
| $N$ | Num. of end hosts | $2^{20}$ (1M) |
| $T$ | Num. of end hosts per rack switch | 32 |
| $S$ | Num. of rack switches | $2^{15}$ (32K) |
| $F$ | Avg. new flows per second per end host | 20 [28] |
| $D$ | Avg. duration of a flow in the flow table | 60 seconds |
| $c$ | Size of counters in bytes | 24 [2] |
| $r_{stat}$ | Rate of gathering statistics | 1-per-second |
| $p$ | Num. of bytes in a packet | 1500 |
| $f_m$ | Fraction of mice | 0.99 |
| $f_e$ | Fraction of elephants | 0.01 |
| $r_{sample}$ | Rate of sampling | 1-in-1000 |
| $h_{sample}$ | Size of packet sample (bytes) | 60 |

TABLE I: Parameters and typical values for the analytical evaluation

# Datacenter Traffic Classification

- **Simulation results:**

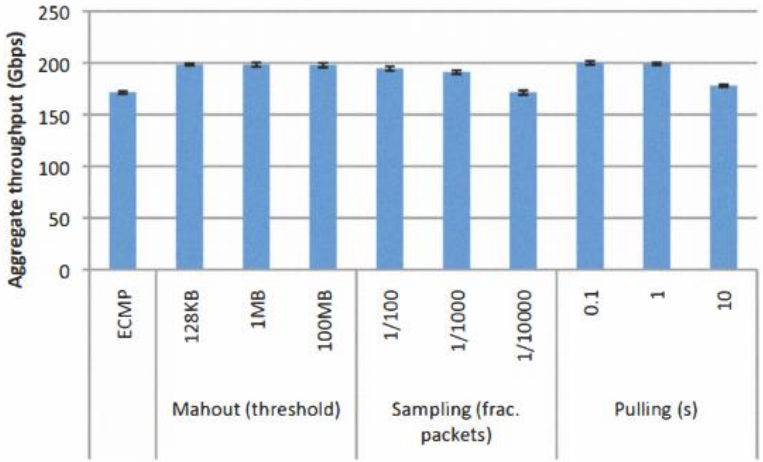



Fig. 4: Throughput results for the schedulers with various parameters. Error bars on all charts show 95% confidence intervals.

| Threshold | 100KB | 200KB | 500KB | 1MB |
|---|---|---|---|---|
| Mahout | 1.531ms | 1.712ms | 3.820ms | 5.479ms |
| Hedera | 189.83ms | 189.83ms | 189.83ms | 189.83ms |

TABLE II: Time it takes to detect an elephant flow at the Mahout controller vs. the Hedera controller, with no other active flows.

# Machine Learning for Traffic Flow Classification

- **Machine Learning Algorithms:**
  - Naïve-Bayes (NBD, NBK)
  - C4.5 Decision Tree
  - Bayesian Network
  - Naïve Bayes Tree

- **Flow and Feature Definitions**
  - Limitation:
    - Packet payload independent
    - Transport layer independent
    - Context limited to a single flow (i.e. no feature spanning multiple flows)
    - Simple to compute

Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(5): 5-16.

# Machine Learning for Traffic Flow Classification

- **Feature Candidates**
  - Protocol
  - Flow duration
  - Flow volume in bytes and packets
  - Packet length (minimum, mean, maximum and standard deviation)
  - Inter-arrival time between packets (minimum, mean, maximum and standard deviation).
- **Feature Reduction**
  - Use CFS and CON to choose features:

| | |
|---|---|
| CFS subset | fpackets, maxfpktl, minfpktl, meanfpktl, stdbpktl, minbpktl, protocol |
| CON subset | fpackets, maxfpktl, meanbpktl, maxbpktl, minfiat, maxfiat, minbiat, maxbiat, duration |

# Machine Learning for Traffic Flow Classification

- **Simulation results**

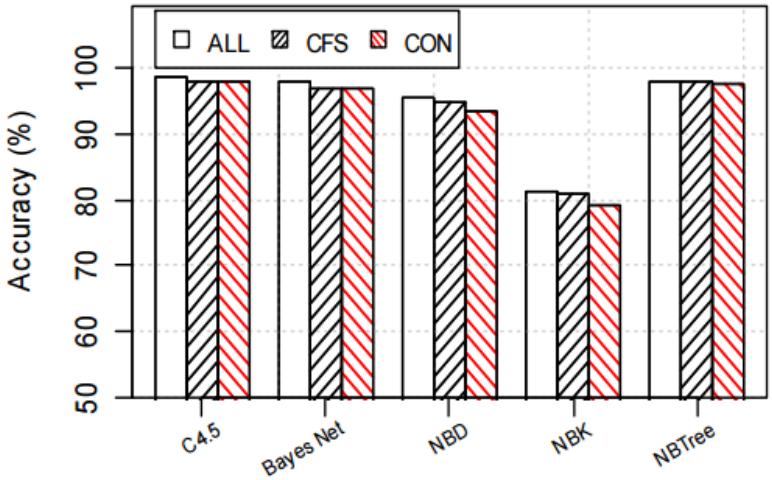



**Figure 2: Accuracy of algorithms using CFS subset, CON Subset and All features.**

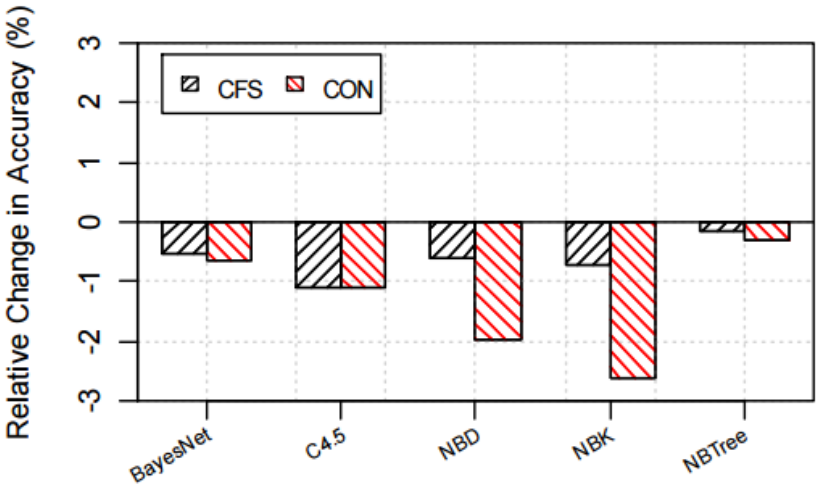



**Figure 3: Relative change in accuracy depending on feature selection metric for each algorithm compared to using full feature set**

# Machine Learning for Traffic Flow Classification

- **Simulation results**

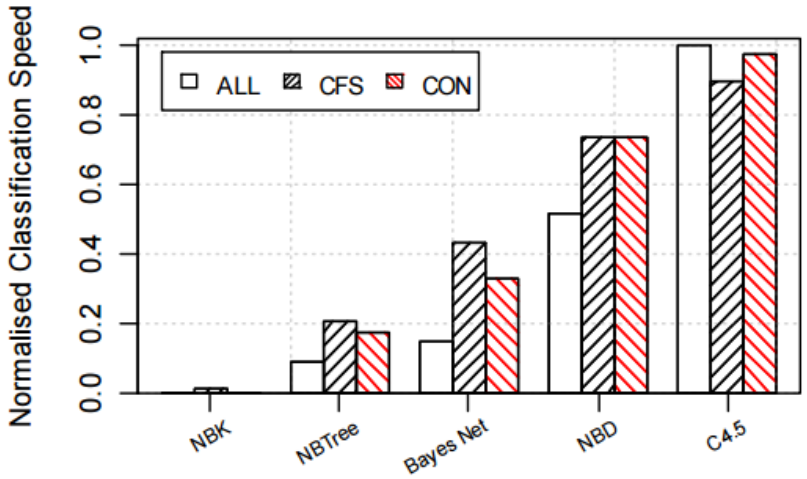



**Figure 4: Normalised classification speed of the algorithms for each feature set**

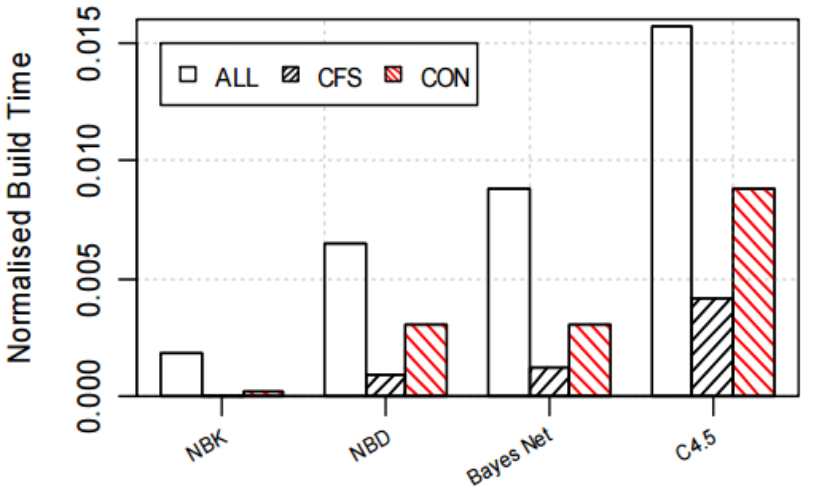



**Figure 6: Normalised build time for each algorithm and feature set except NBTree**


UCDAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu