A Novel Bandwidth Allocation Scheme for OTSS-enabled Flex-grid Intra-datacenter Networks

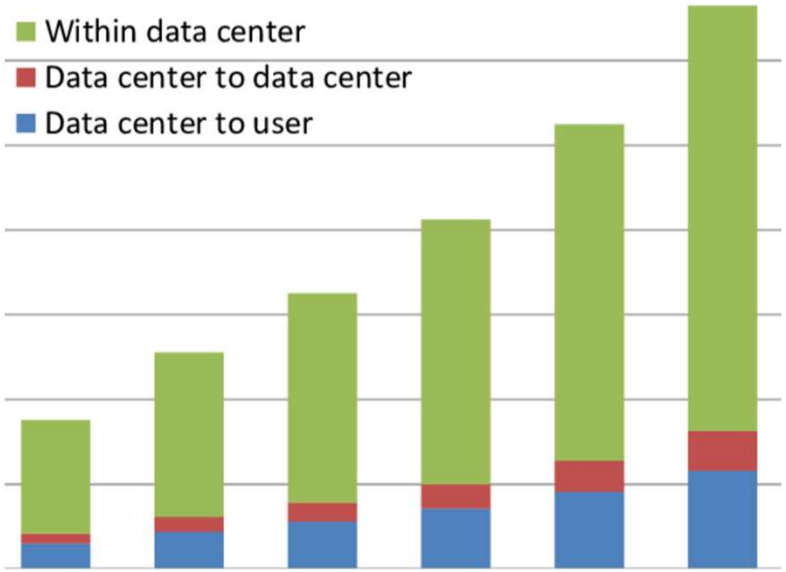
Speaker: Lin Wang

Research Advisor: Biswanath Mukherjee

UCDAVIS

Motivation

- **Traffic demand increasing in datacenter networks**
  - Cloud-service, parallel-computing, etc., lead to huge amount of intra datacenter traffic growth.
  - Cisco forecasts 31% increase per year of datacenter traffic by 2021

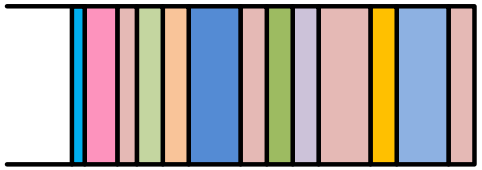


Datacenter traffic loads is growing

Introduction

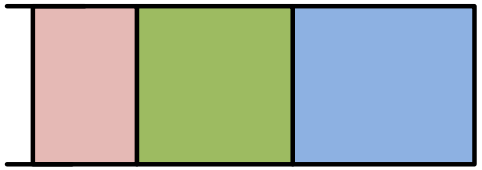
- **Datacenter traffic measurements**
- A large fraction of datacenter traffic is carried in a small fraction of flows.
- 90% of flows carry less than 1MB of data, called " mice flows".
- More than 90% of bytes are transferred in flows greater than 100MB, called "elephant flows"

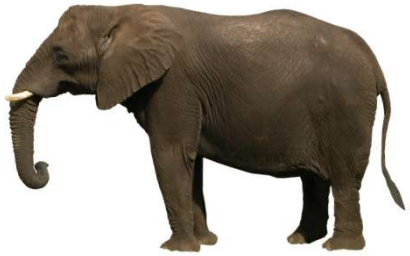
Mice VS. Elephant Flow



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

transactional traffic, web browsing, search queries (≈ 90% traffic)





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

bulk data transfer, data backup, virtual machine migration (≈ 10% traffic)

Elephant flow detection

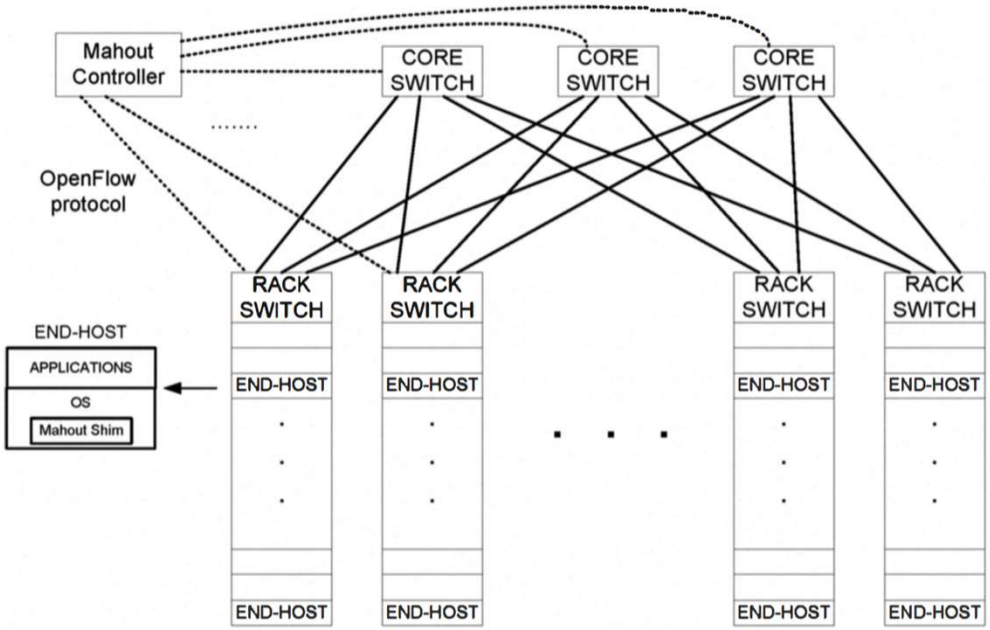
- **Application identify elephant flows**
- Impractical for traffic management in datacenter, as each application needs to be modified to support.

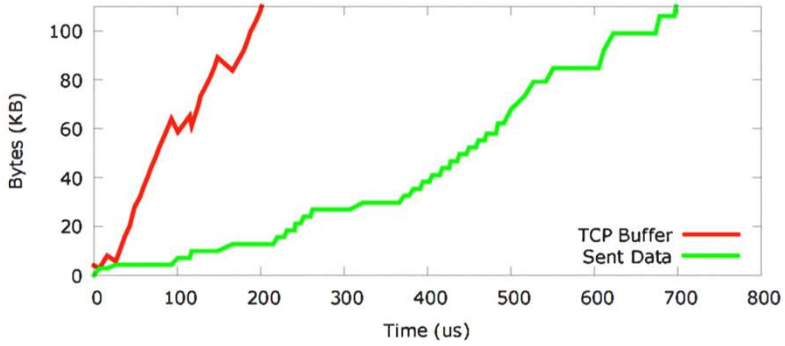
- **Maintain per-flow statistics**
- Not scale to large datacenter networks

- **Sampling**
- Is not reliable to detect an elephant flow before it has carried more than 10K packets, roughly 15MB.

Elephant flow detection

- **End host monitor**
- Monitor flows at origin end hosts. When detect an elephant flow, it marks subsequent packets of that flow using in-band signaling mechanism.





Amount of data observed in the TCP buffers vs. data observed at the network layer for a flow.

Curtis, Andrew R., Wonho Kim, and Praveen Yalagandula. "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection." In INFOCOM, Proceedings IEEE, pp. 1629-1637. IEEE, 2011.

Needs for a transparent fine-grained optical network

- **Optical networks: enormous transmission bandwidth.**
  - High-order modulation (PAM4): increase per-channel capacity.
  - space-division-multiplexing: increase spatial channels.

- **Mismatch between application demands and optical channel capacity.**
  - Traffic grooming is the first proposal.
  - Drawback of grooming: energy, latency, security, etc.

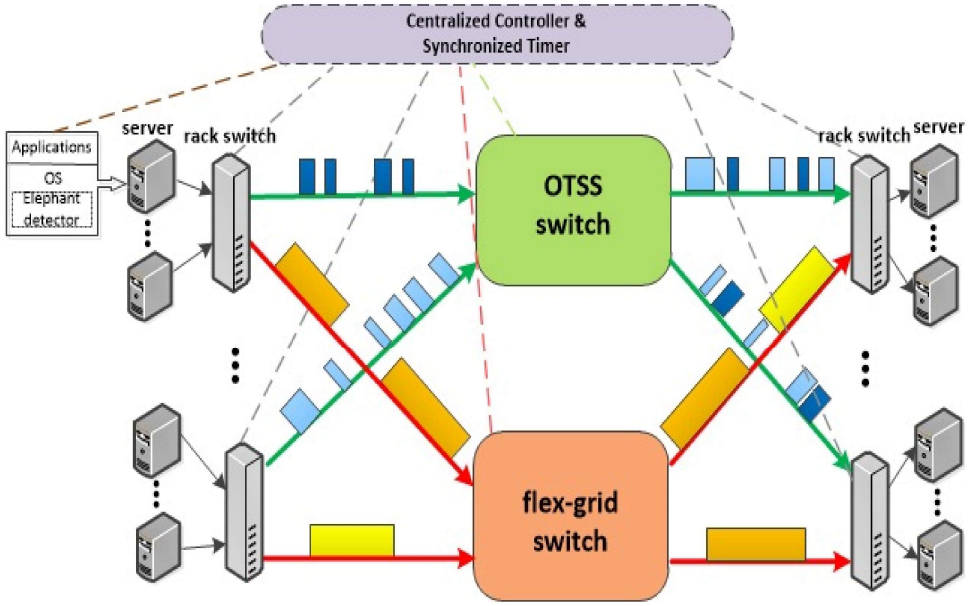
Build a transparent, bufferless, fine-grained, WDM-like network

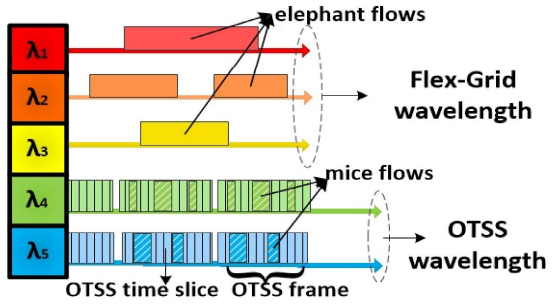
- Why WDM can avoid collision?
  - Wavelength channels are separated by a global coordinate.
  - (frequency! All the same in different nodes)
- Time synchronization: a global coordinate in temporal domain
  - Definite time, all nodes are synchronized for a global coordinate.
- Temporally-statistical multiplexing for asynchronous transmission based on synchronized global time.
  - We call it: Optical Time Slice Switching (OTSS).

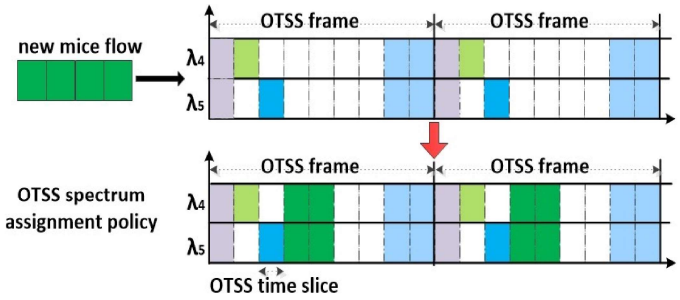
OTSS Principle

Control message format

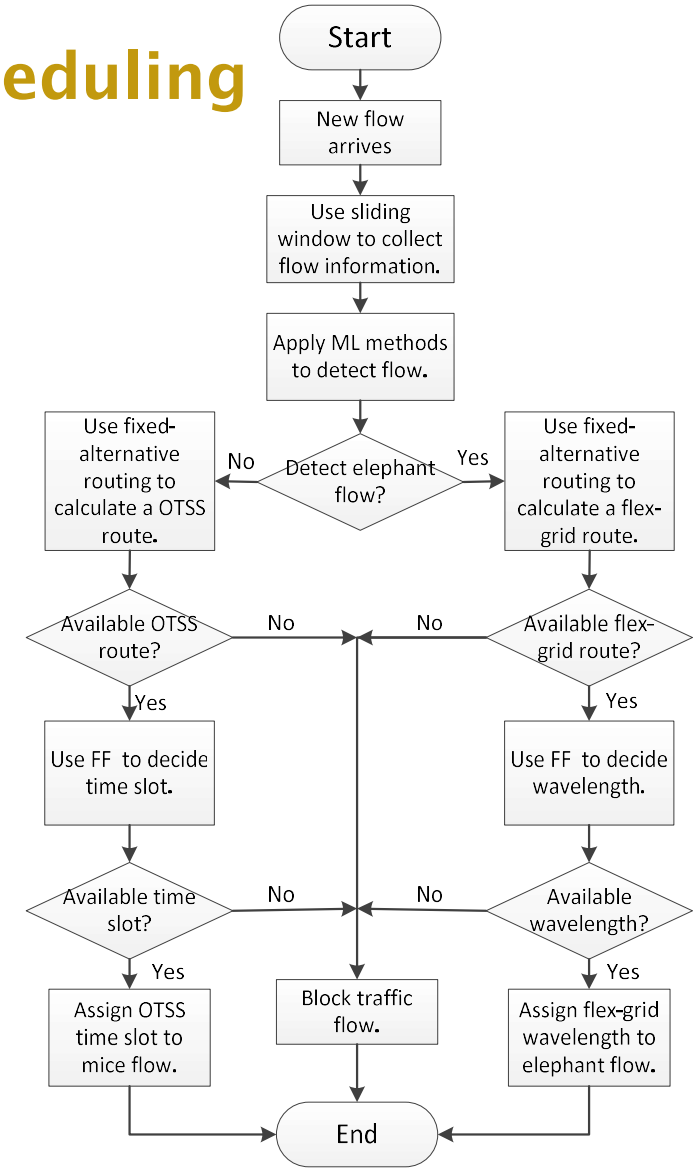
Frame Length	Operation Time	Operation Type
Fiber ID	Wavelength ID	
Traffic ID	Traffic Source	Traffic Destination



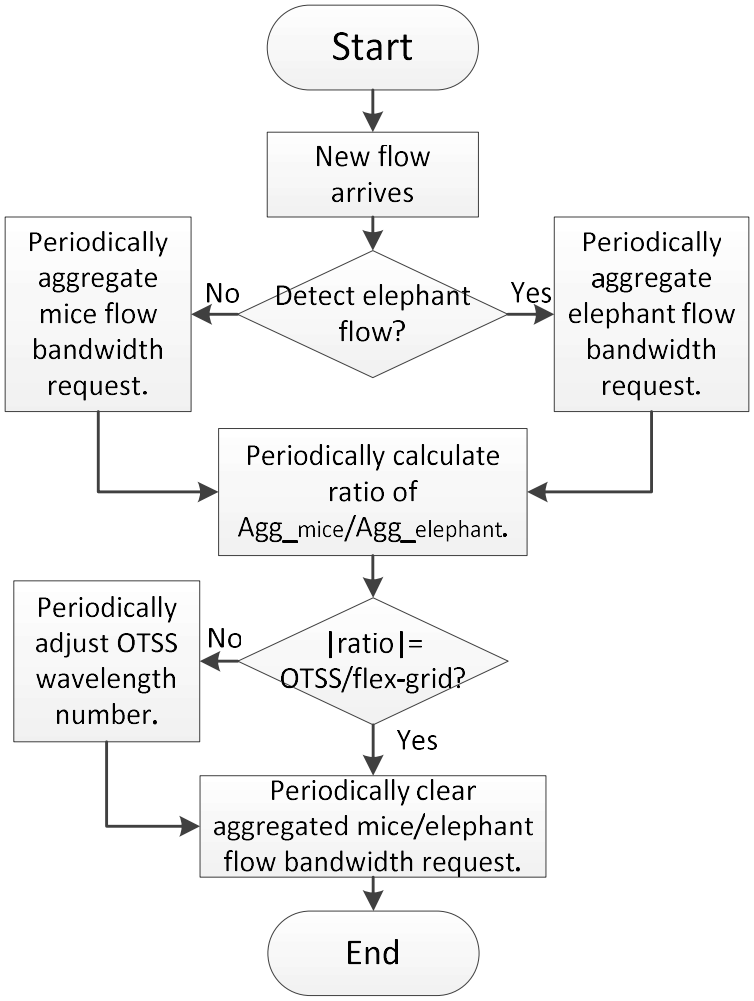




Static Flow scheduling



Dynamic Flow scheduling



UCDAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu