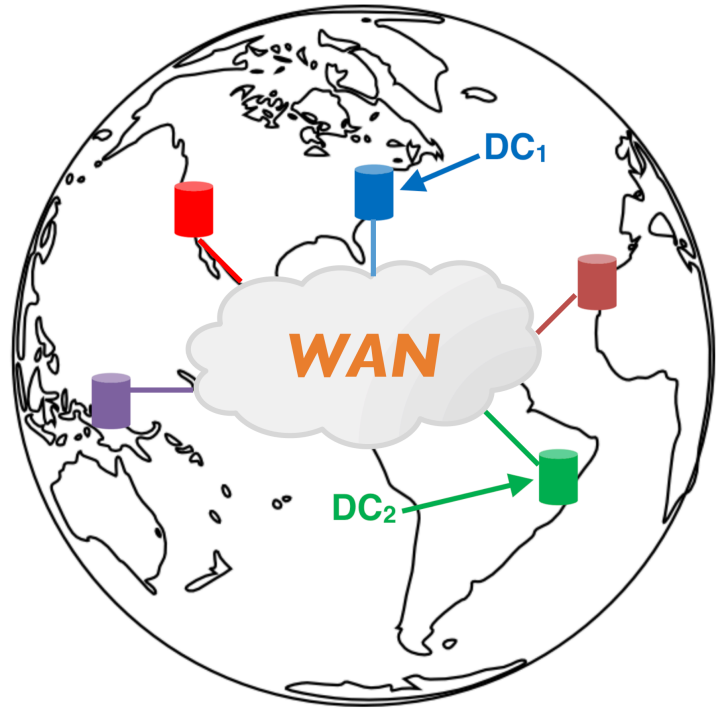
To Relay or Not to Relay for Inter-Cloud Transfers?

Fan Lai, Mosharaf Chowdhury, Harsha Madhyastha

Background

- **Over 40 Data Centers (DCs) on EC2, Azure, Google Cloud**
  - A geographically denser set of DCs across clouds
- **Cloud apps host on multiple DCs**
  - Web search, Interactive Multimedia
  - Low latency access, privacy regulations
- **Massive data across geo-distributed DCs**

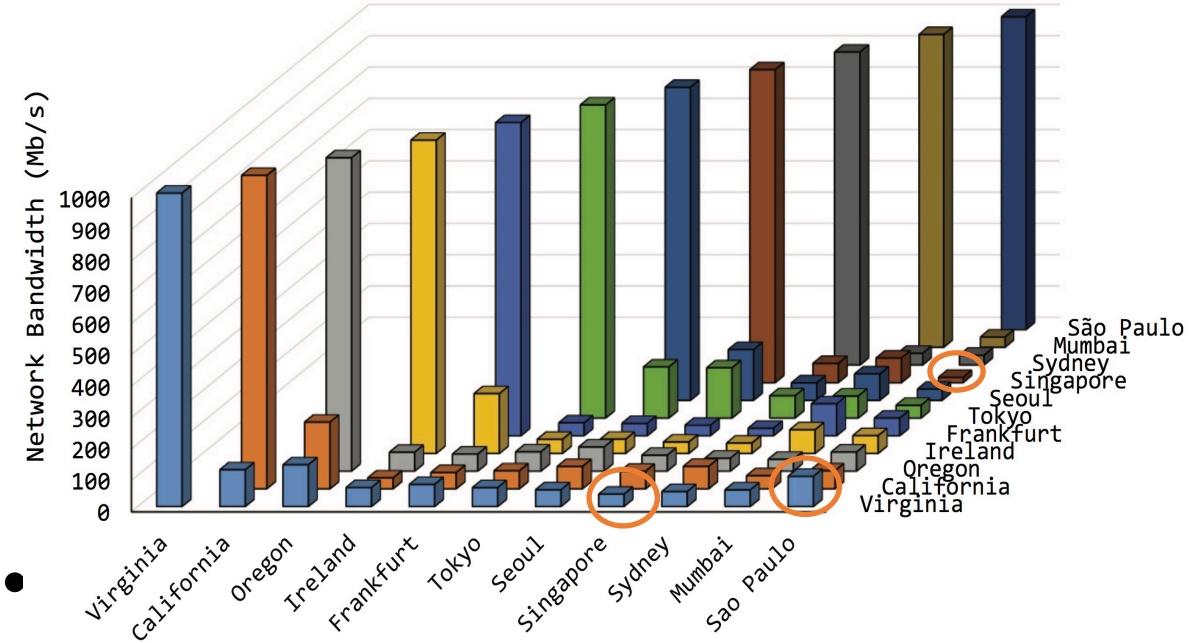


WAN is Crucial for Geo-distributed Service

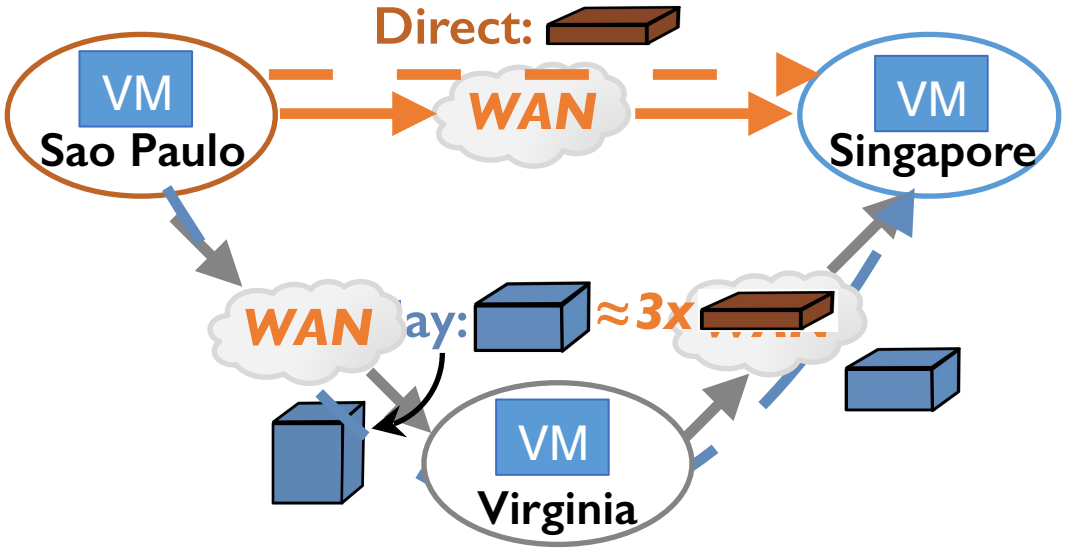
- **Bandwidth-intensive transfers**
  - **Geo-distributed replication**: Web search, cloud storage
  - **Inter-DC Routing**: SWAN[SIGCOMM'13], Pretium[SIGCOMM'16], etc
  - **Big data analytics**: Iridium [SIGCOMM'15], Clarinet [OSDI'16] …
  - …
- **Latency-sensitive traffic**
  - **Interactive service**: Skype, Hangout
  - **Transaction processing**: SPANStore[SOSP'13], Carousel[SIGMOD'18], etc
  - …

Prior Efforts: WAN b/w varies spatially

- **WAN bandwidth(b/w) *varies* *significantly* between different regions**
  - Close regions have more than 12× of the b/w than distant regions[1]



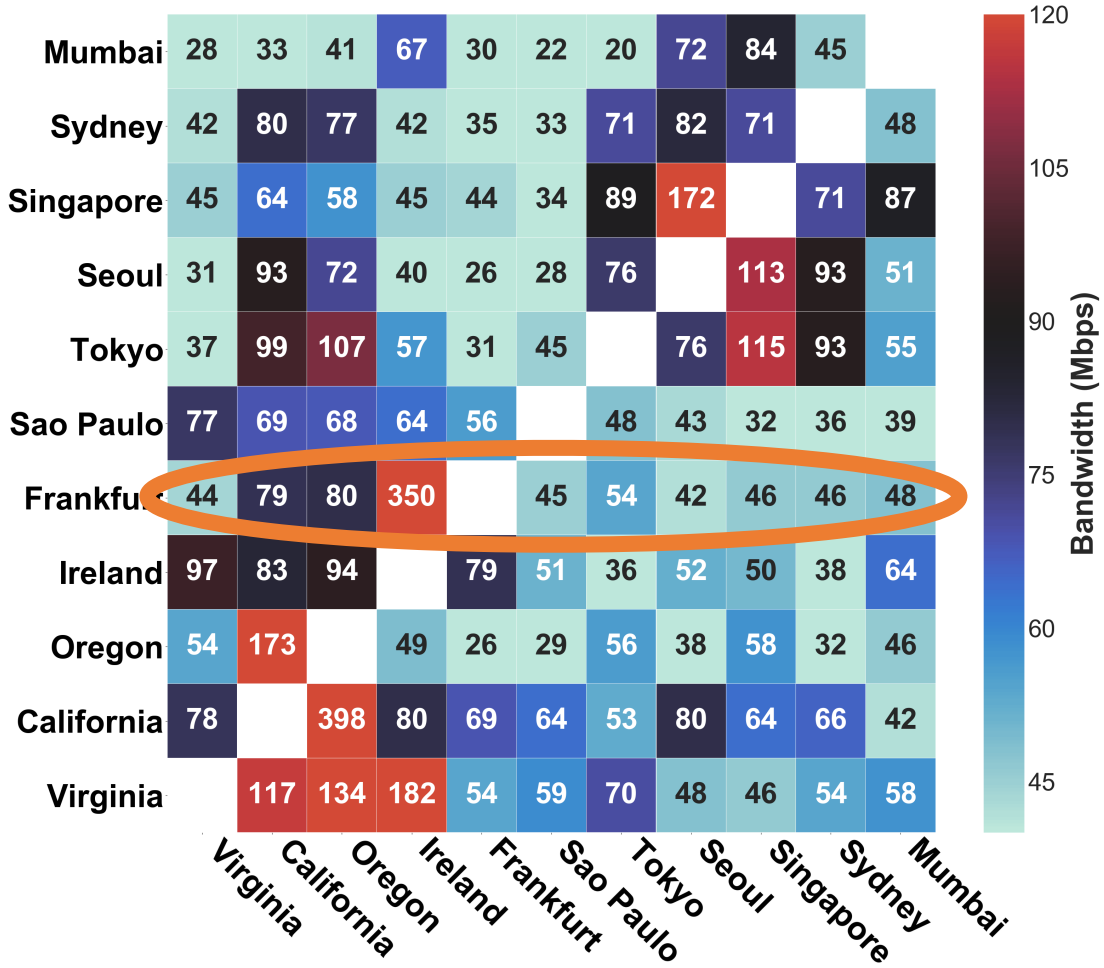
Bandwidth Measurement across 11 EC2 regions[1]



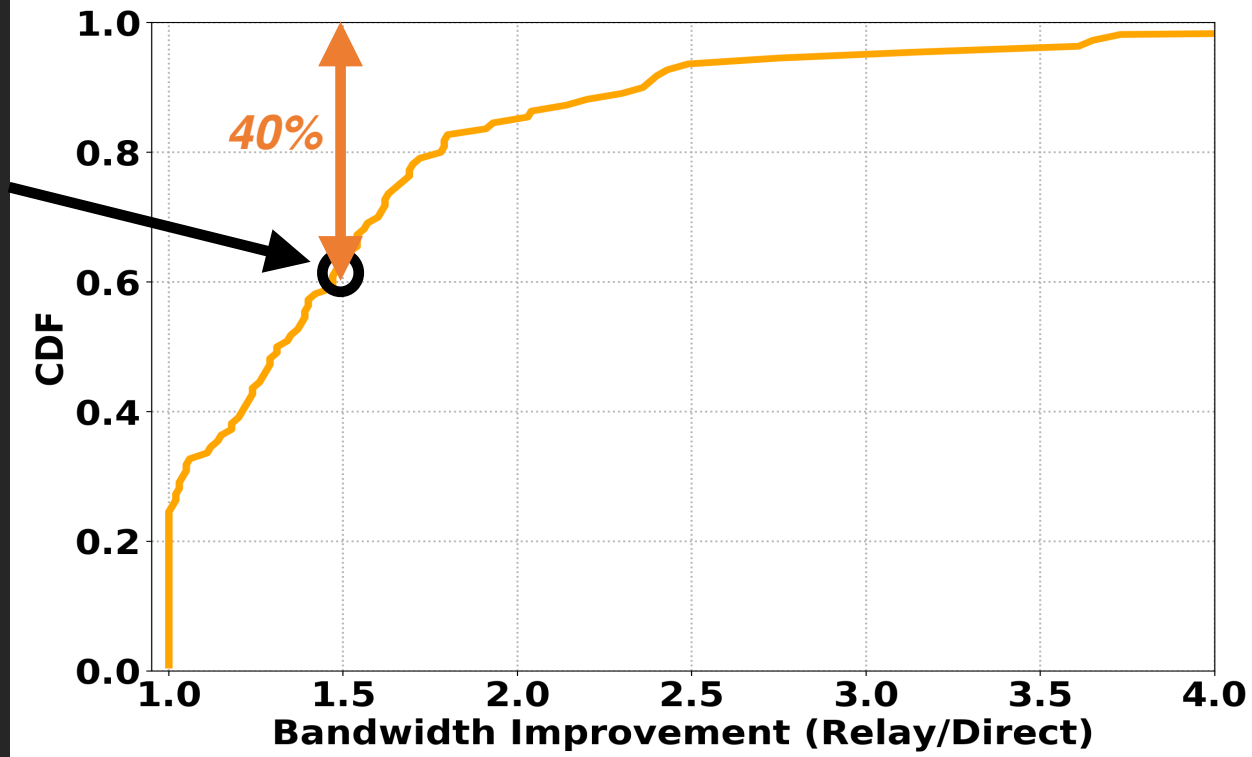
[1] "Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds." NSDI'17

WAN Bandwidth Varies Spatially

- Reproduce prior measurements
  - 11 EC2 regions, 110 inter-DC pairs
  - Tools: *iperf* (TCP)
- Heterogeneous link capacity
  - Varies between the same type of VMs
  - Lower b/w between distant regions
- Relay should work pretty well



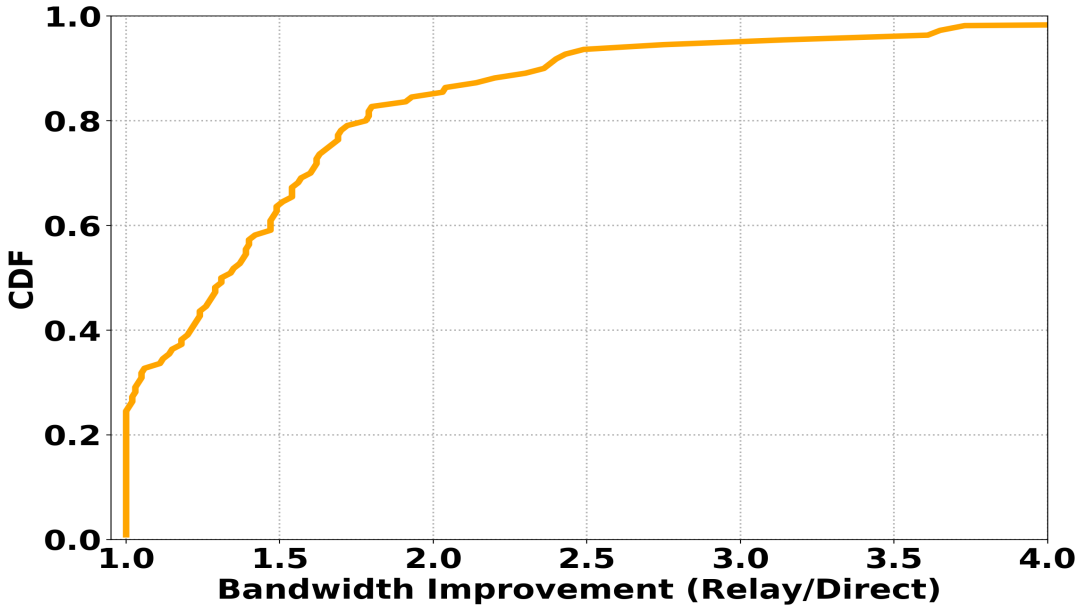
About 40% percent data transfers between EC2 regions can have more than 1.5x bandwidth increase via relay

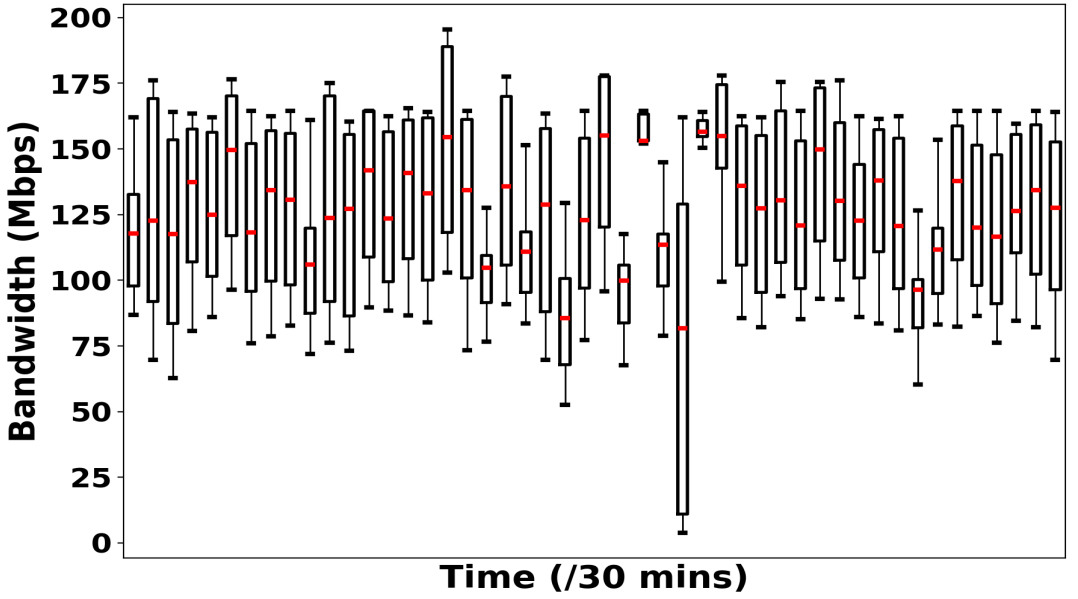


Bandwidth improvement via best relay on EC2

How to identify and tackle this complicated WAN?

- *Heterogeneous across regions*
- *Dynamic runtime environment*
- *Great complexity in sys design*





How to identify and tackle this complicated WAN?

- Heterogeneous across regions
- Dynamic runtime environment
- Great complexity in sys design

Assumptions in prior measurements:

- Default TCP setting works well
- Single TCP is representative enough for the available b/w

What if we Break Down these assumptions ?

- *Default TCP setting works well*
- *Single TCP is representative enough for the available b/w*

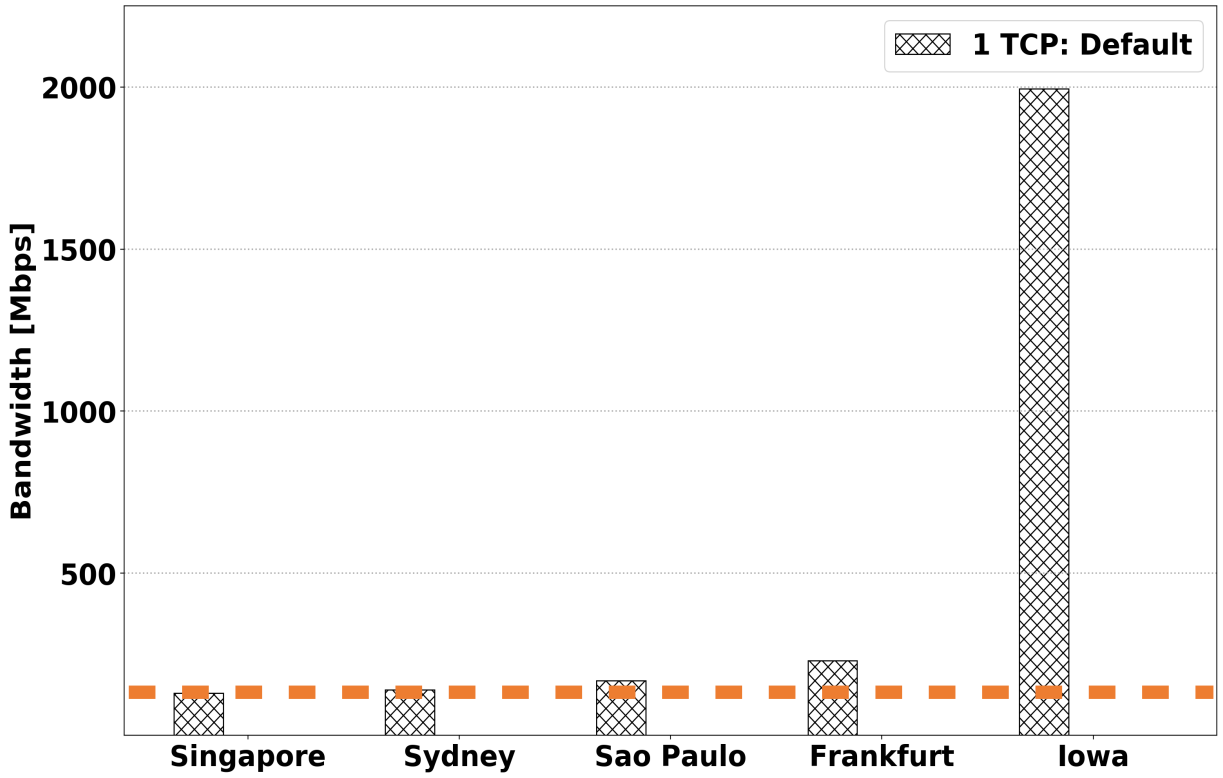
#1:Whether the b/w still varies spatially ?

#2:Whether the b/w still varies temporally?

#3: How much room for WAN improvement via relay?

Default TCP Setting may be Sub-optimal

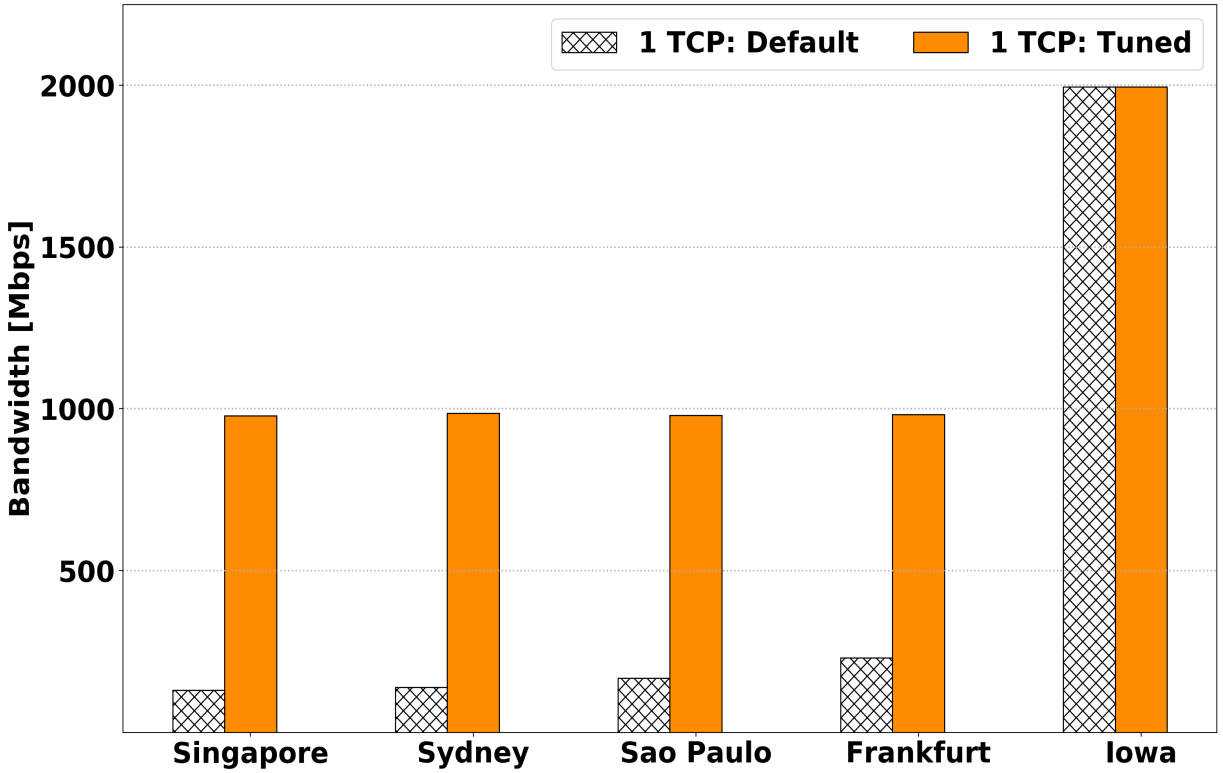
- **B/w varies across regions**
  - Lower b/w between distant regions
  - RTT varies across regions
- **Max TCP window is bounded**
  - TCP throughput is RTT-based
-



Google: Bandwidth to Iowa

Default TCP Setting is Sub-optimal

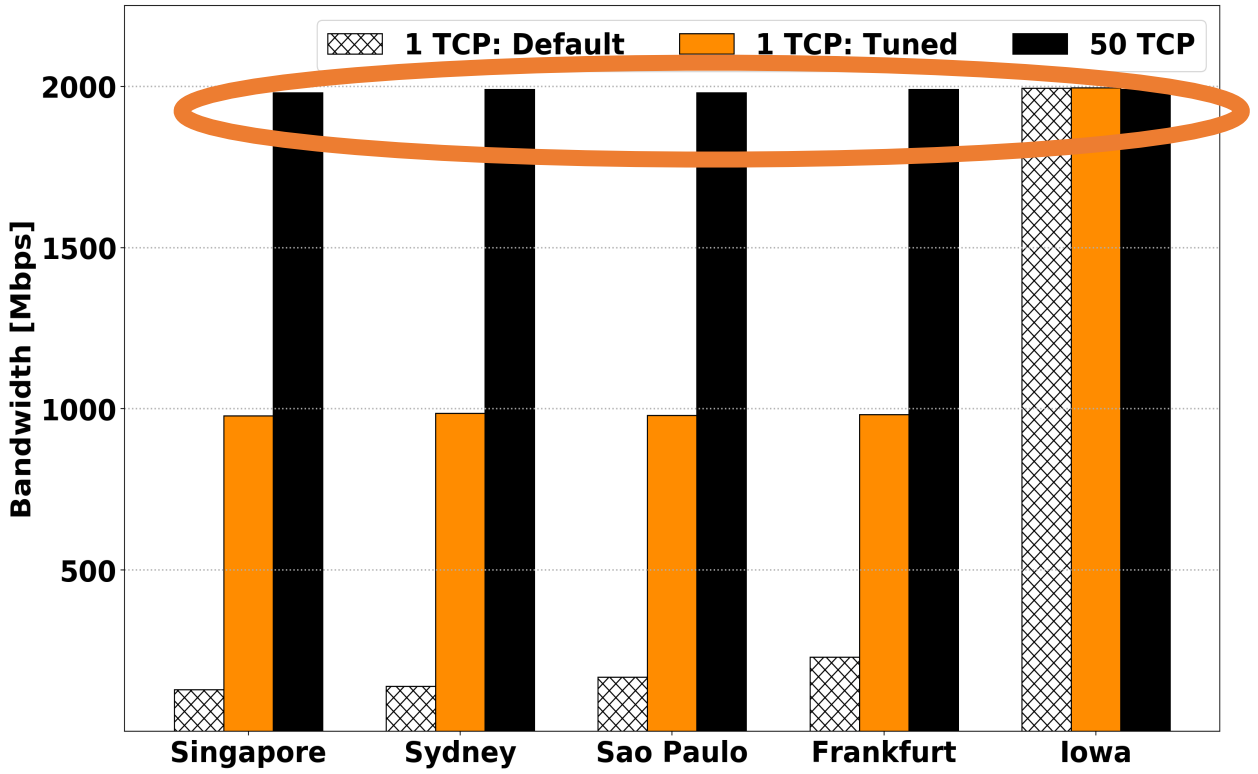
- **B/w varies across regions**
  - Lower b/w between distant regions
  - RTT varies across regions
- **Max TCP window is bounded**
  - TCP throughput is RTT-based
- **Per-TCP rate limit on the WAN**



Google: Bandwidth to Iowa

Single TCP is not Representative

- Single TCP underutilize the b/w
  - Use multiple TCPs
- Per-VM cap for outbound rate
  - Per-TCP rate limit < Per-VM cap
- Aggregate b/w is homogeneous
  - VM-cap works on all connections



Google: Bandwidth to Iowa

What if we Break Down these assumptions ?

- ~~Default TCP setting works well~~
- ~~Single TCP is representative enough for the available b/w~~

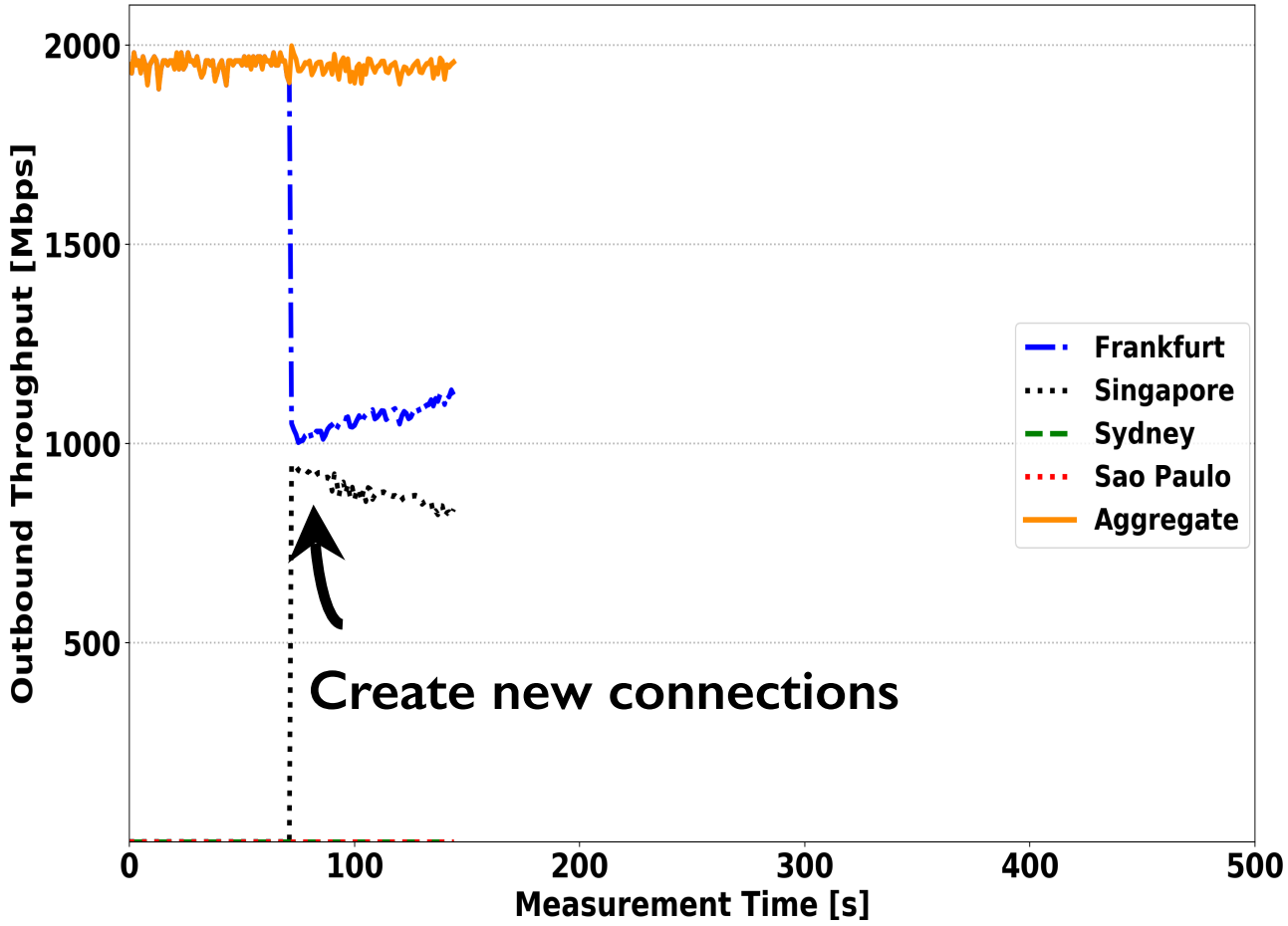
#1:Whether the b/w still varies spatially ? Often Homogeneous

#2:Whether the b/w still varies temporally?

#3: How much room for WAN improvement via relay?

Available B/w is often Stable

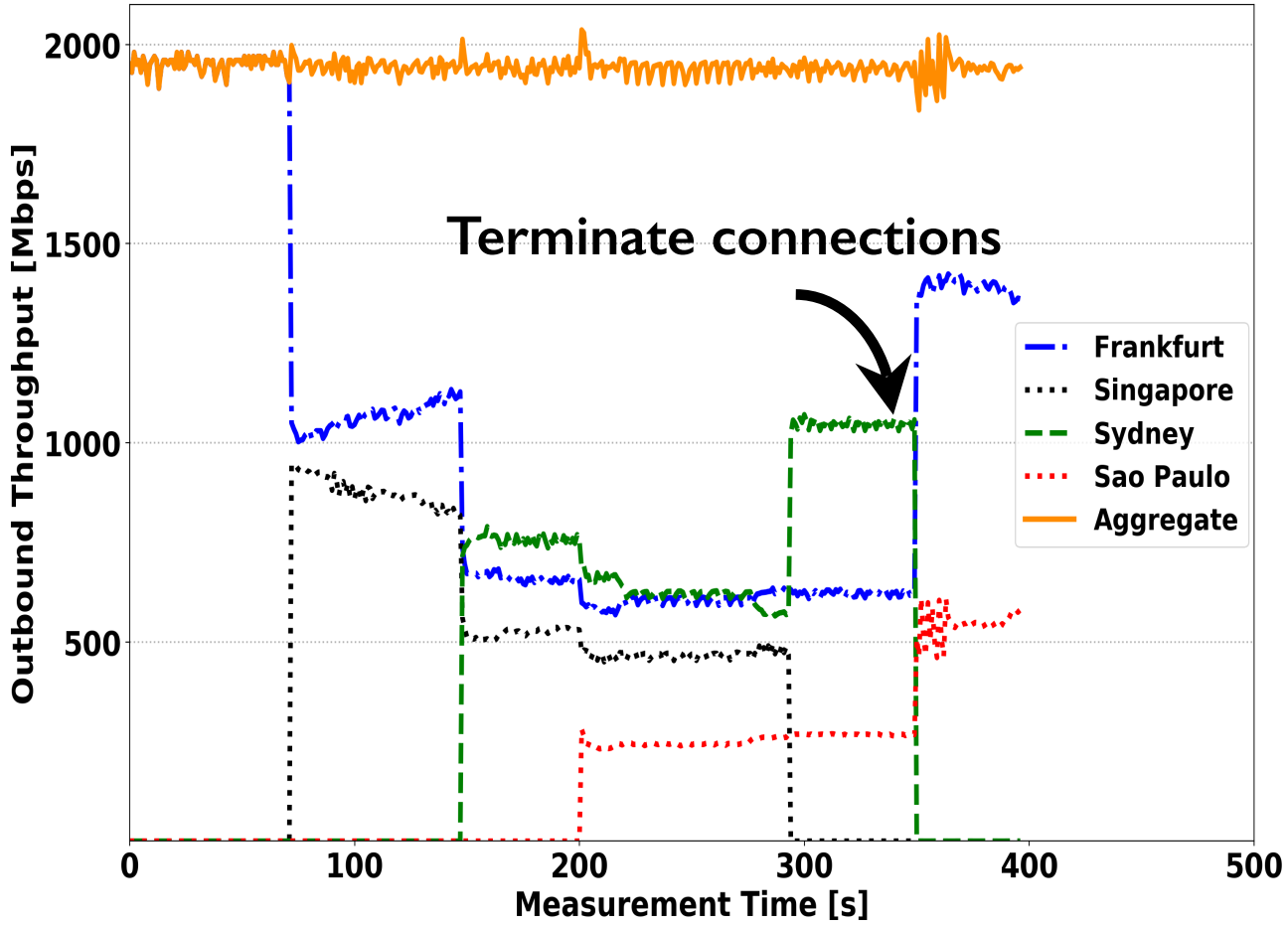
- Measurement setup
  - Create/terminate connections
- Inter-DC connections share the VM-cap
-



Google: Throughput from Iowa

Available B/w is often Stable

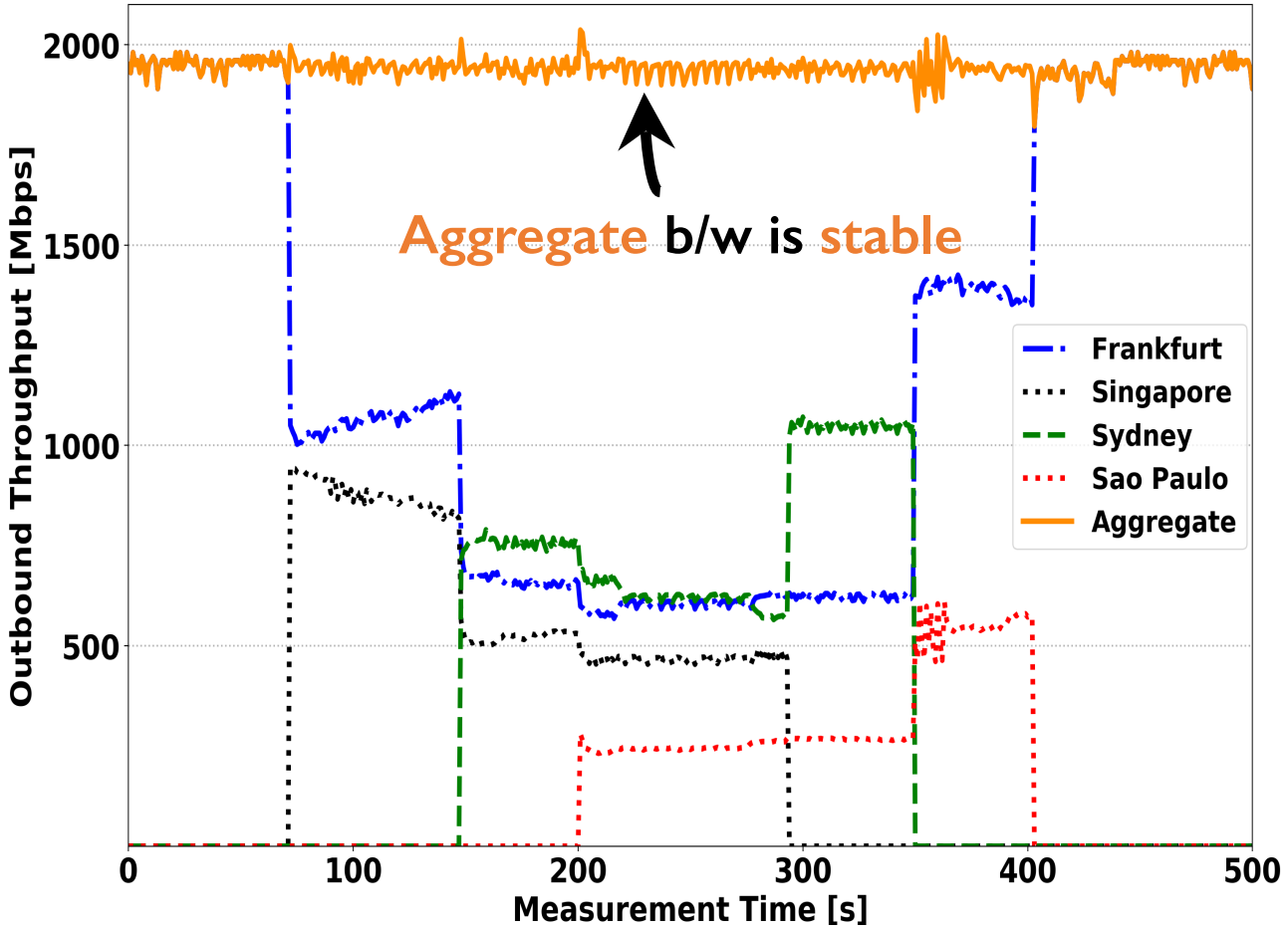
- Measurement setup
  - Create/terminate connections
- Inter-DC connections share the VM-cap
-



Google: Throughput from Iowa

Available B/w is often Stable

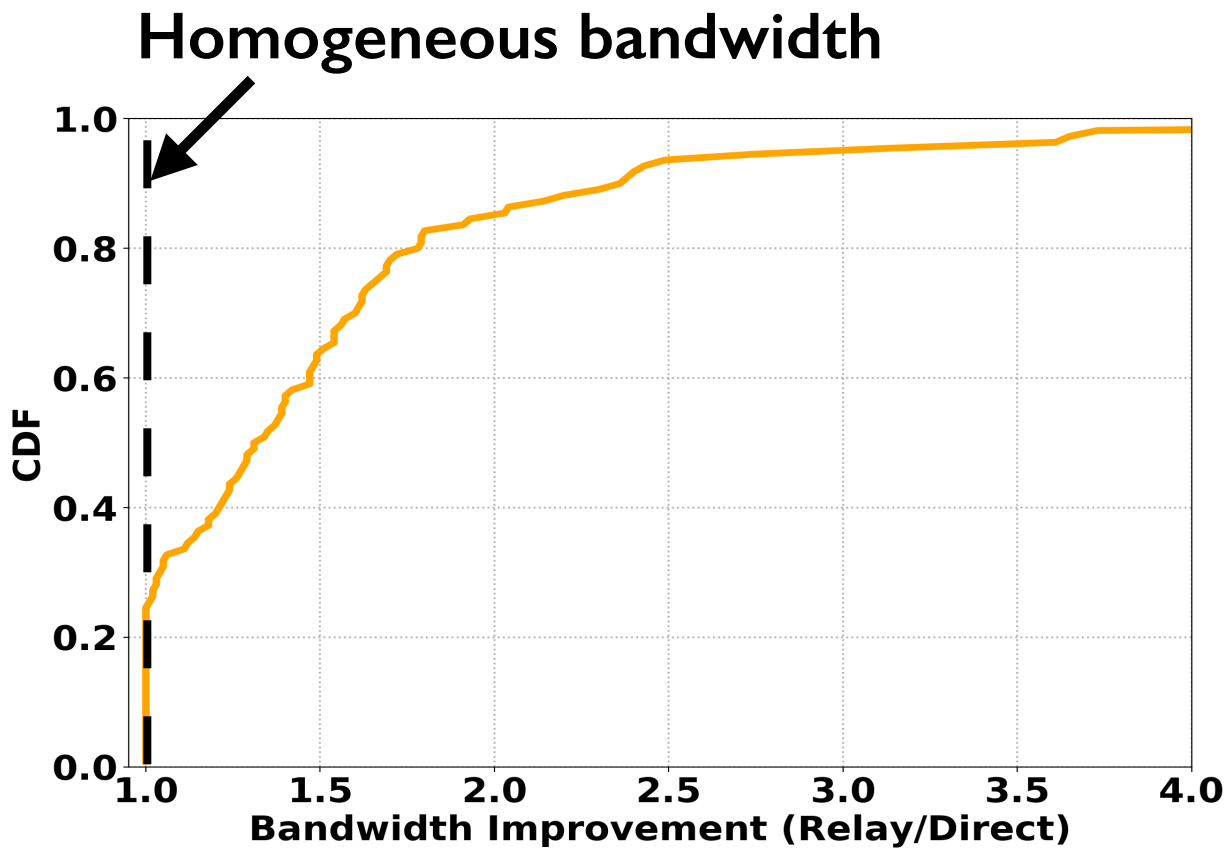
- Measurement setup
  - Create/terminate connections
- Inter-DC connections share the VM-cap
- *Max* b/w (VM cap) is stable



Google: Throughput from Iowa

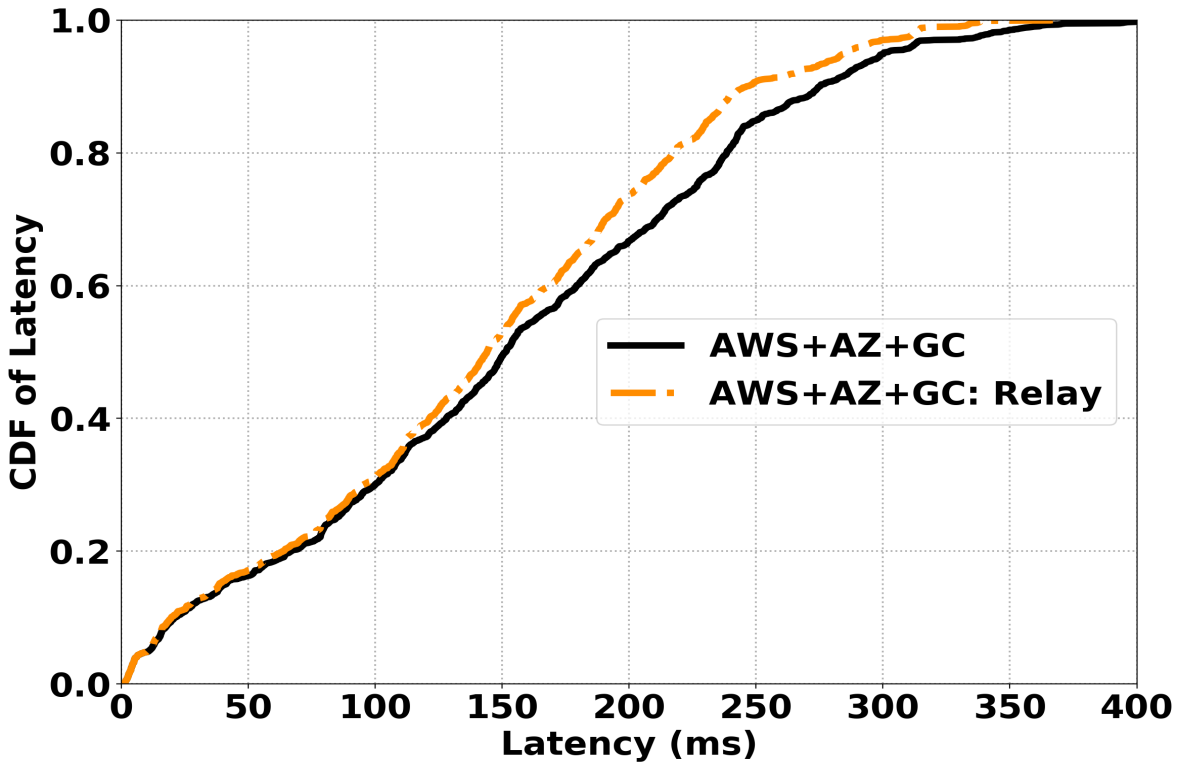
*Maximum available **bandwidth***

- *Homogeneous across regions*
- *Stable over time*
- *Varies with VM instances*
- *Performance can be predict-able w/o great sys complexity*

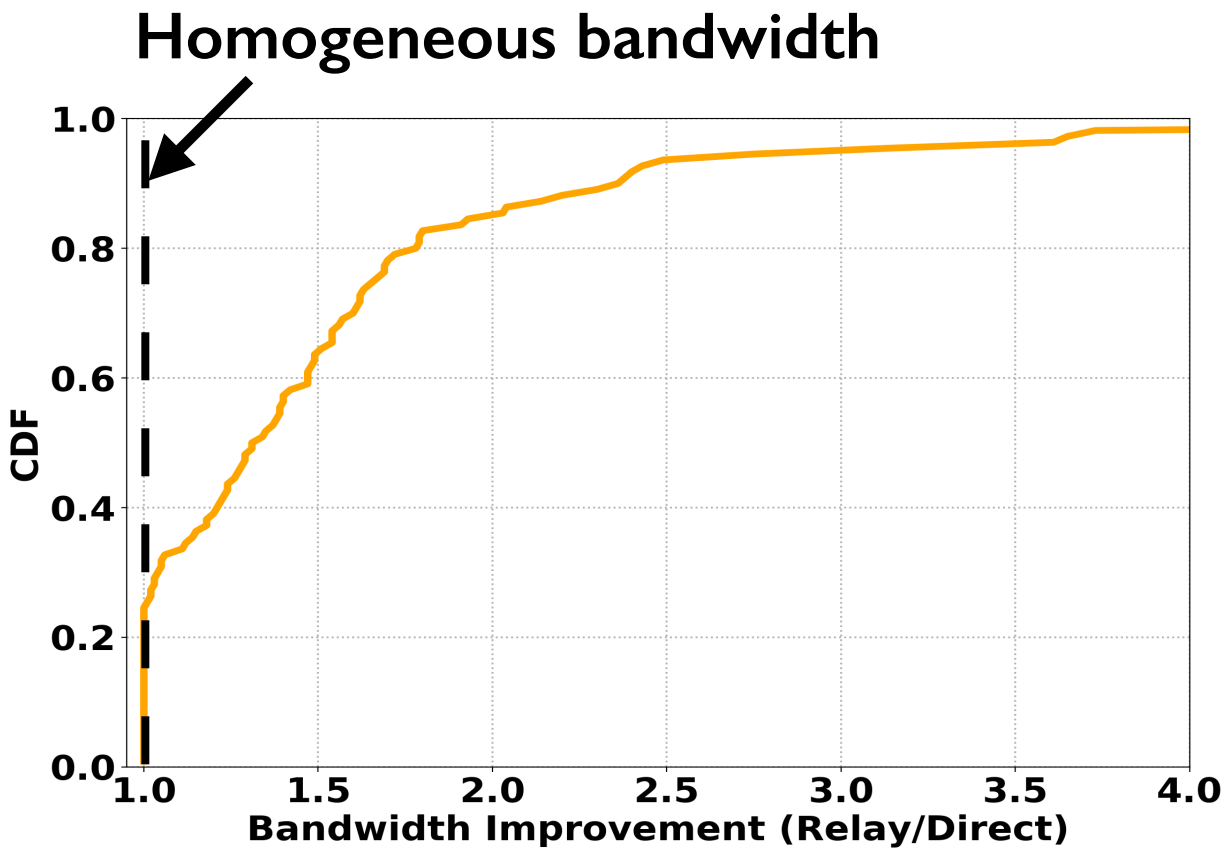


What will happen if the b/w is homogeneous ?

Little Scope for Optimization via Inter-DC Relay



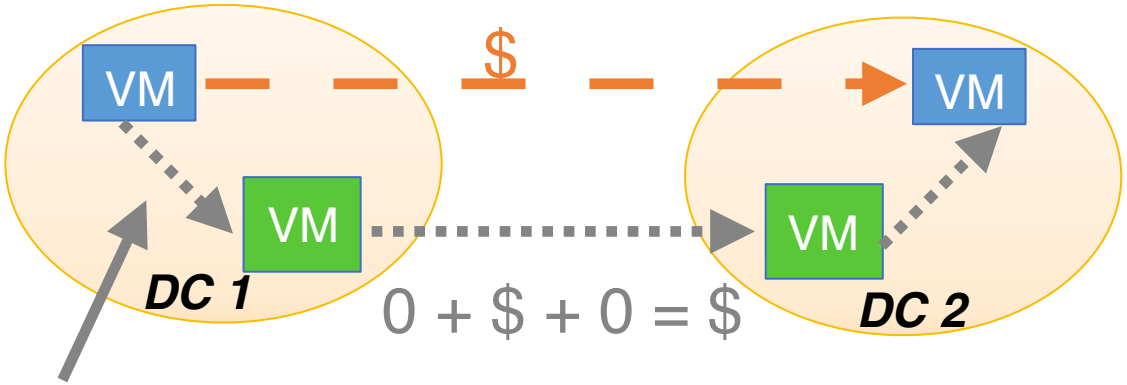
Latency Measurement across 40 DCs



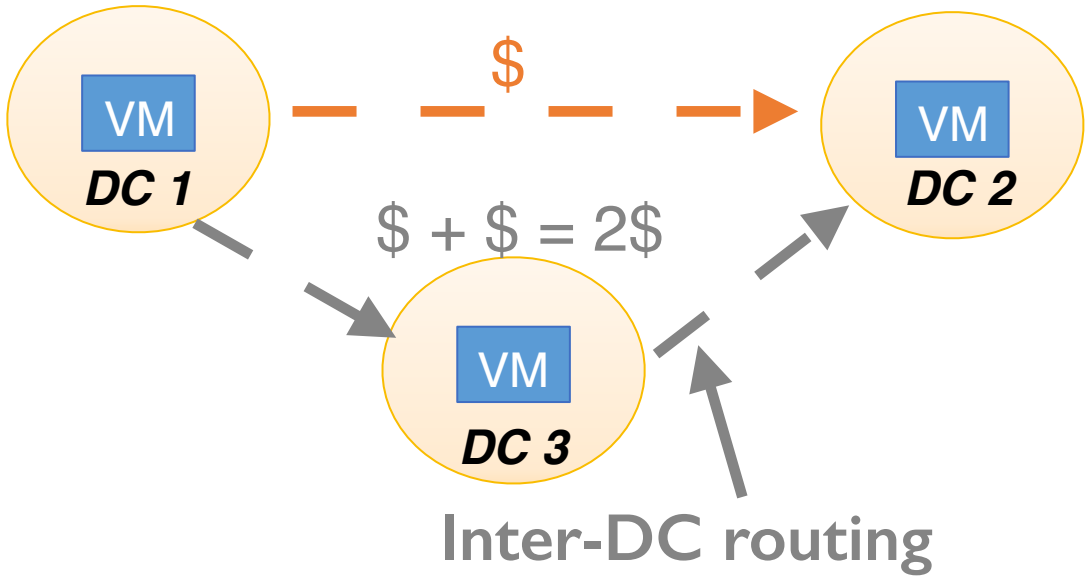
What will happen if the b/w is homogeneous ?

Takeaway

- Intra-DC relay from poor performance VMs to high performance VMs
  - Gain more inter-DC bandwidth without extra costs for transfers
  - Routing through a third DC takes your money away



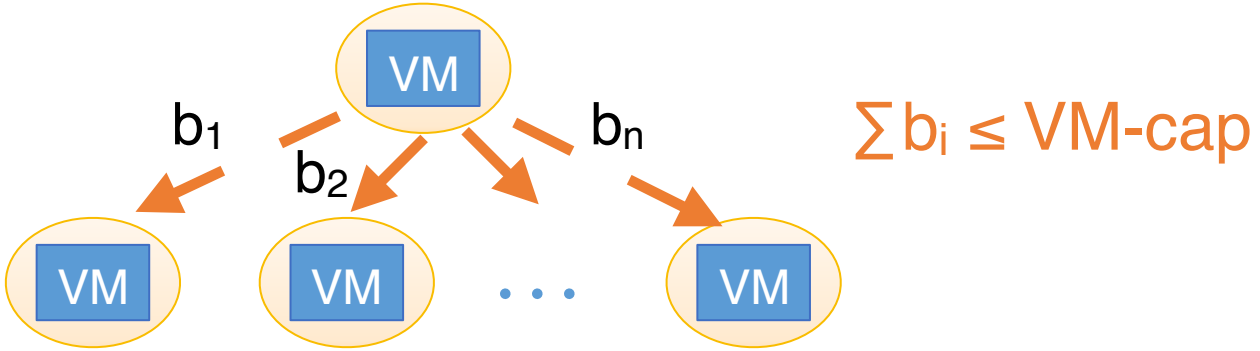
Intra-DC relay



Inter-DC routing

Takeaway

- **Turn to the optimization of bandwidth contentions inside VMs**
  - VM-cap VS link-level optimizations used in existing GDA work
  - VM-aware VS WAN-aware
- **Bandwidth measurements are far from complete**
  - More than 40 VM instance types
  -



Thank you!

Questions?

fanlai@umich.edu

#1:Whether the b/w still varies spatially ? Often Homogeneous

#2:Whether the b/w still varies temporally? Often Stable

#3: How much room for WAN improvement via relay?

Case by case