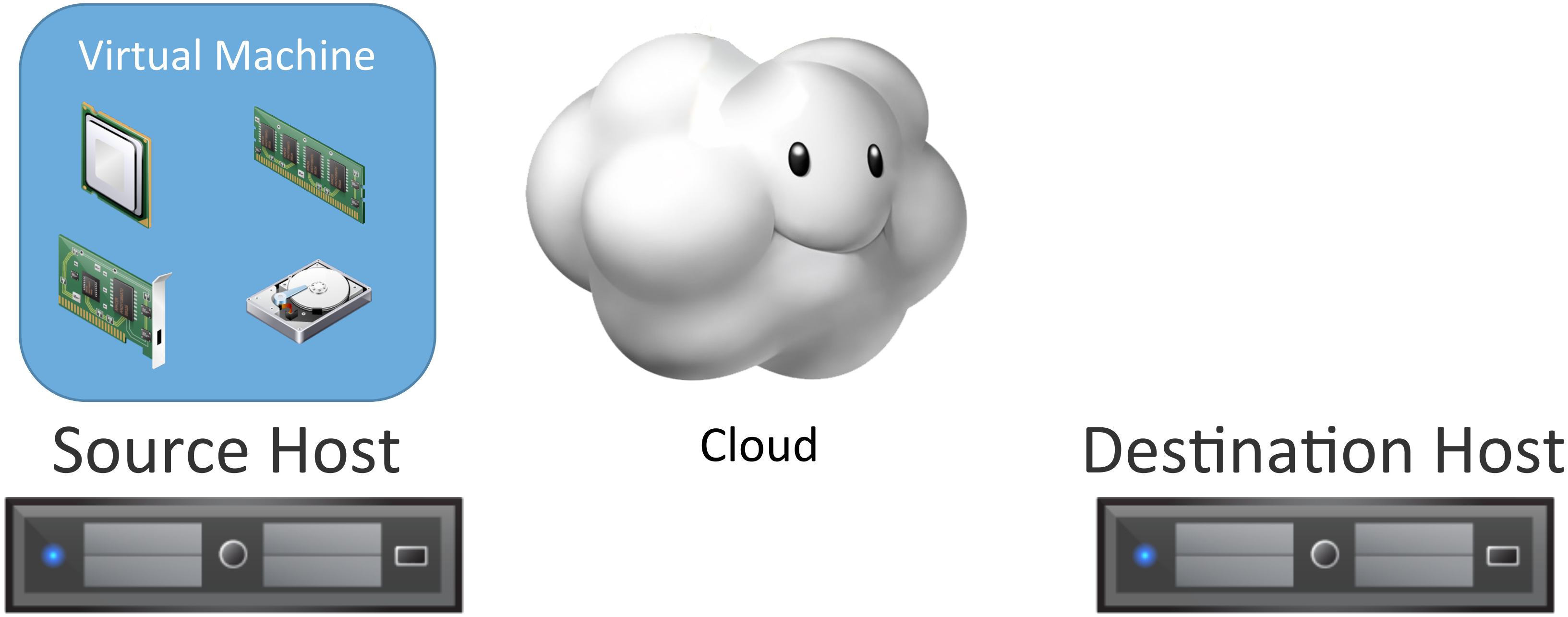
XvMotion: Unified Virtual Machine Migration over Long Distance

Ali José Mashtizadeh, Min Cai, Gabriel Tarasuk-Levin, Ricardo Koller, Tal Garfinkel, Sreekanth Setty

Stanford University – VMware, Inc.

Live Migration



Virtual Machine

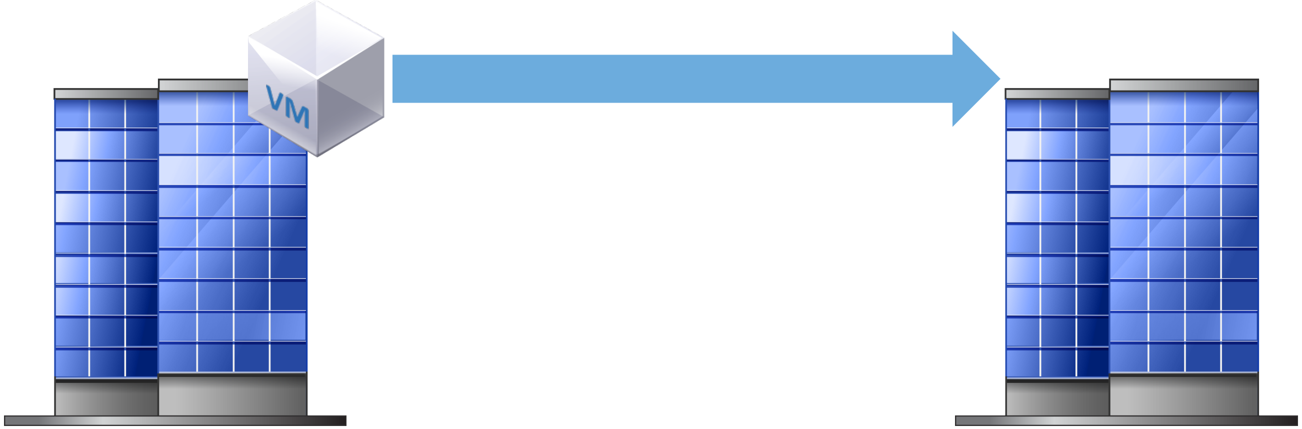
Source Host

Cloud

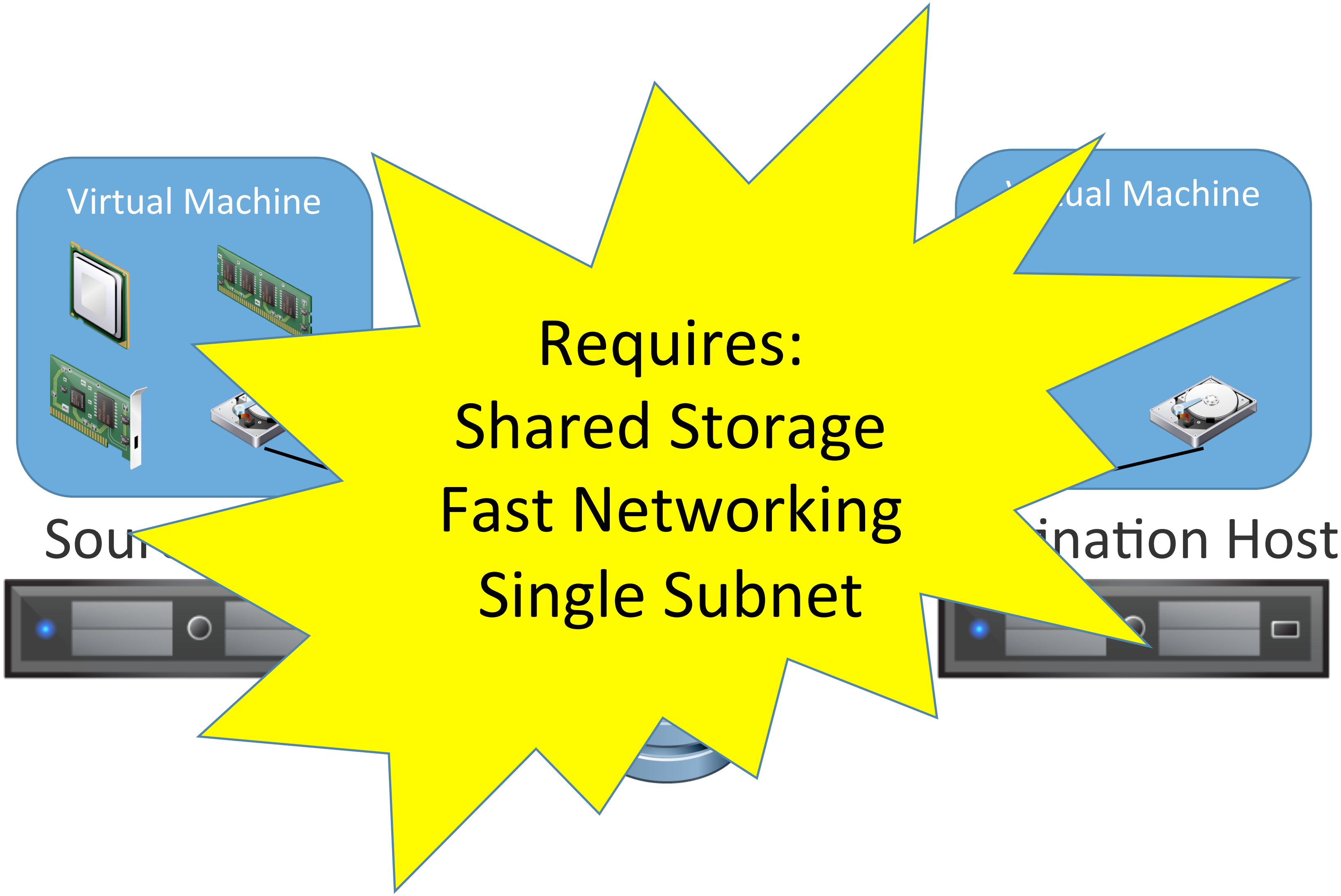
Destination Host

Migration Benefits

- Test to production migrations
- Disaster Preparedness Testing
- Cross-Datacenter Load Balancing
- Shared-Nothing Architecture



Migration in Practice



Requires:
Shared Storage
Fast Networking
Single Subnet

This is not what we really want

- Migrations are limited to machines that:
  With shared storage, fast networks, and same LAN

- Technological Changes:
  - Shared nothing application architectures
  - Network mobility possible: LISP, OTV, VXlan, OpenFlow (SDN)
  - Very large virtualized datacenters

- No reason for these limitations any longer
- Customers have new use cases

XvMotion

- XvMotion: First commercially viable WAN migration

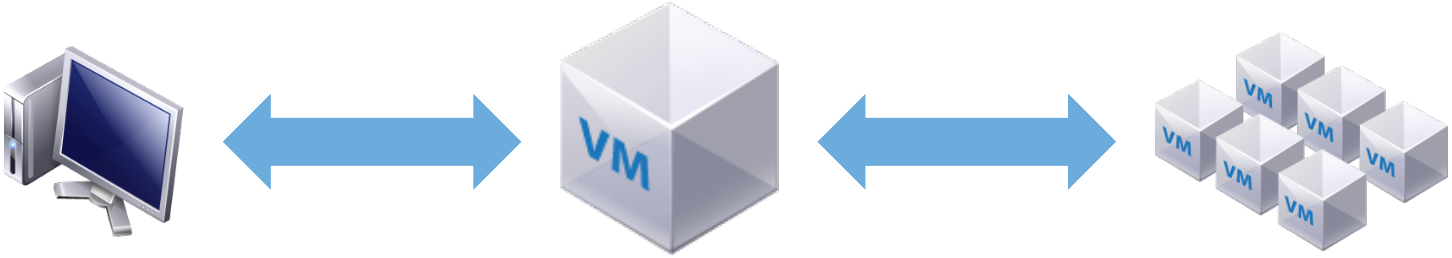
- Achieve Low Downtime AND Atomic Switchover
- Uses Asynchronous IO Mirroring

- Principle:
  Flow control mechanisms for memory and disk

Customer Scenario

	LAN	WAN
Bandwidth	10 Gbps (sometimes 40 Gbps)	1 Gbps or less
Latency	<1 ms	~100 ms
Typical Network	Dedicated NIC(s)	Shared connection between sites

Example Workload



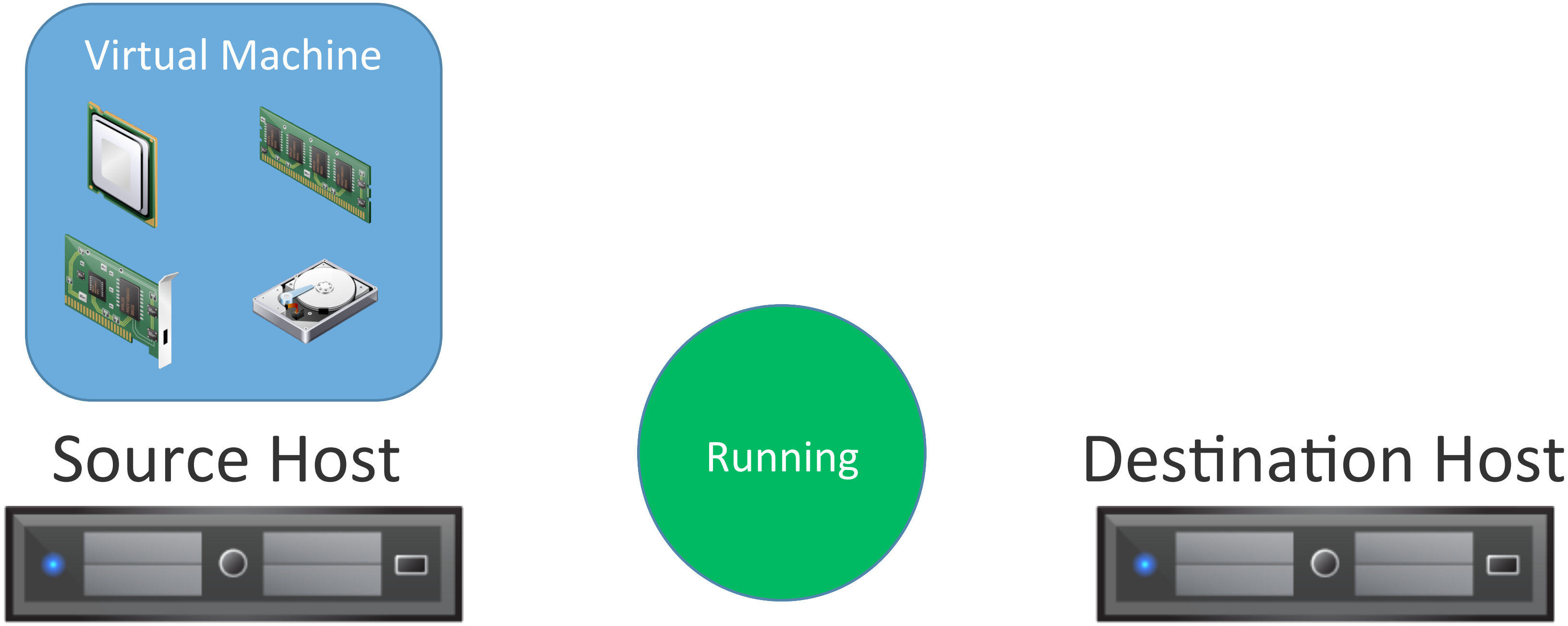
Client

Application

Highly Available
Database

HA Timeouts several seconds
TCP Timeouts 120 seconds

Downtime (Switchover Time)



Virtual Machine

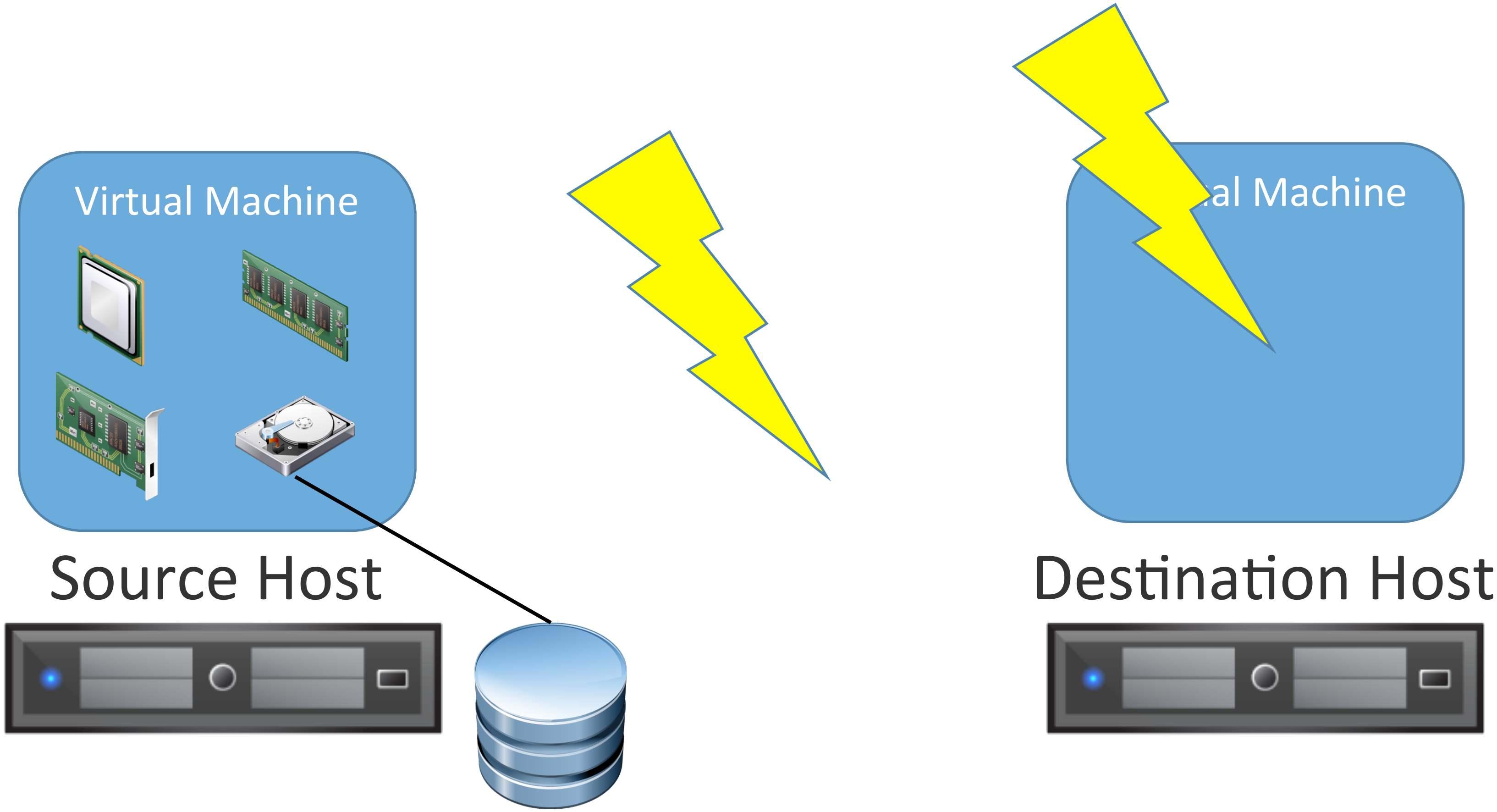
Source Host

Running

Destination Host

Goal: Less than 1 Second of Downtime

Atomicity



Goal: Atomic switchover

XvMotion

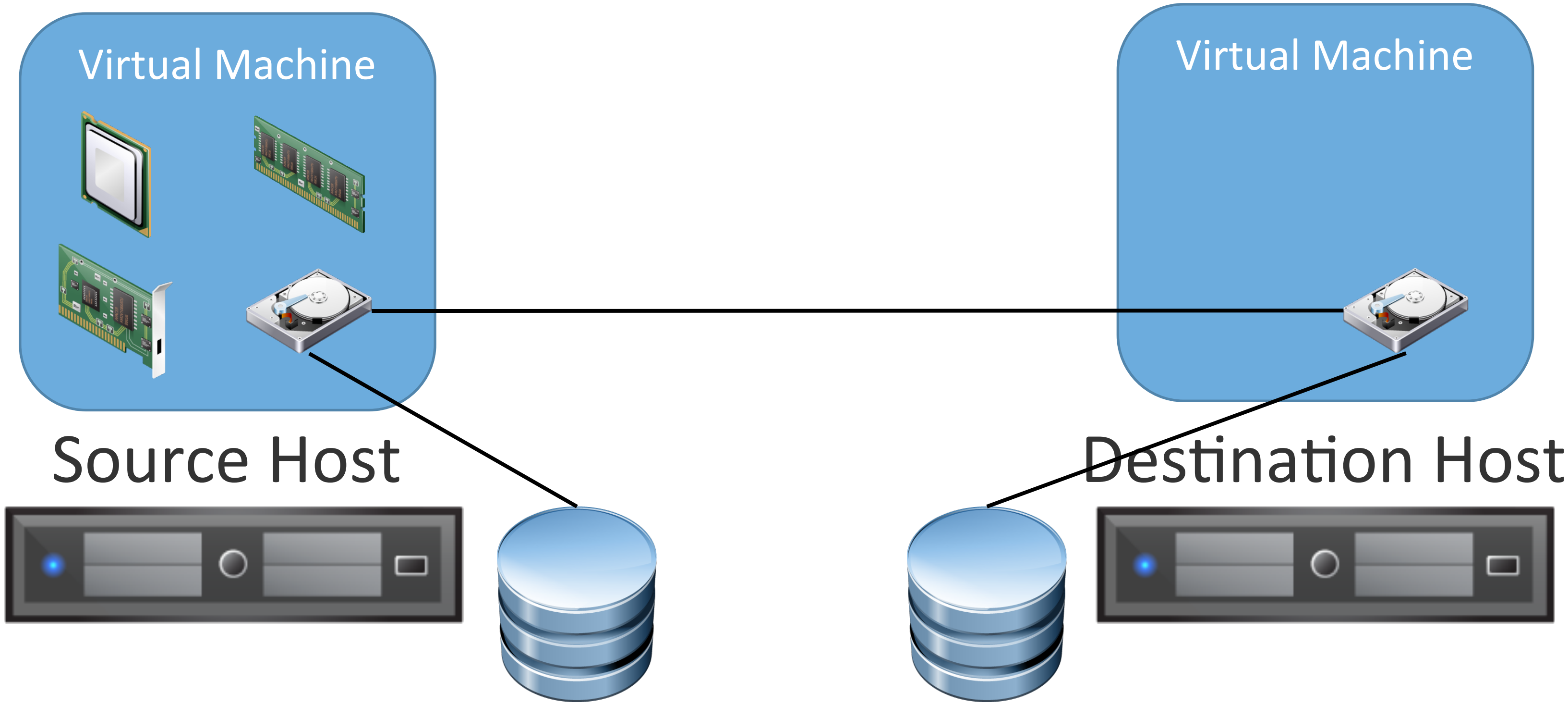
- Unifies Memory Migration and Storage Migration
  - Tolerates wide area network bandwidth/latency and reliability
  - Tolerates heterogeneous storage performance
  - Downtimes and workload impact comparable to local migration
  - Atomic switchover

- Deployed in customer metro area networks
- Cross continent migration e.g. Palo Alto to India is practical

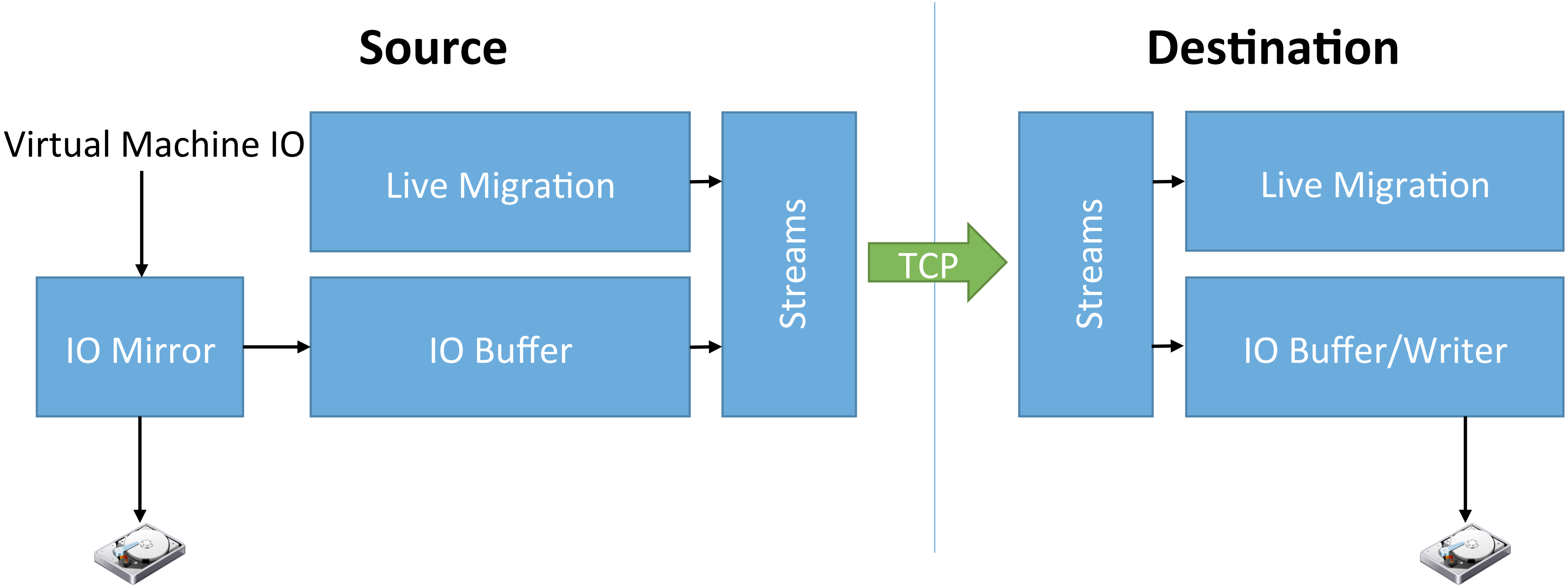
Overview

- **Architecture Overview**
- Wide Area Memory Migration
- Wide Area Storage Migration
- Evaluation

Unified Live Migration



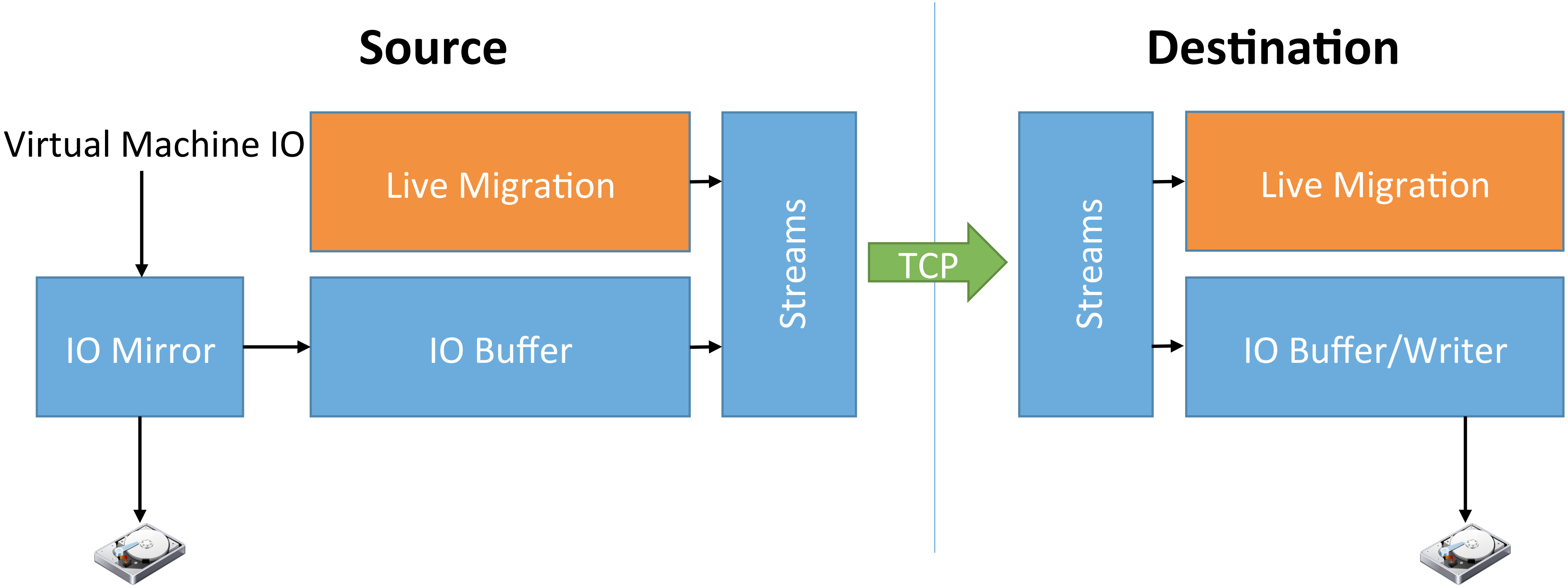
XvMotion Architecture



Overview

- Architecture Overview
- **Wide Area Memory Migration**
- Wide Area Storage Migration
- Evaluation

XvMotion Architecture: Live Migration



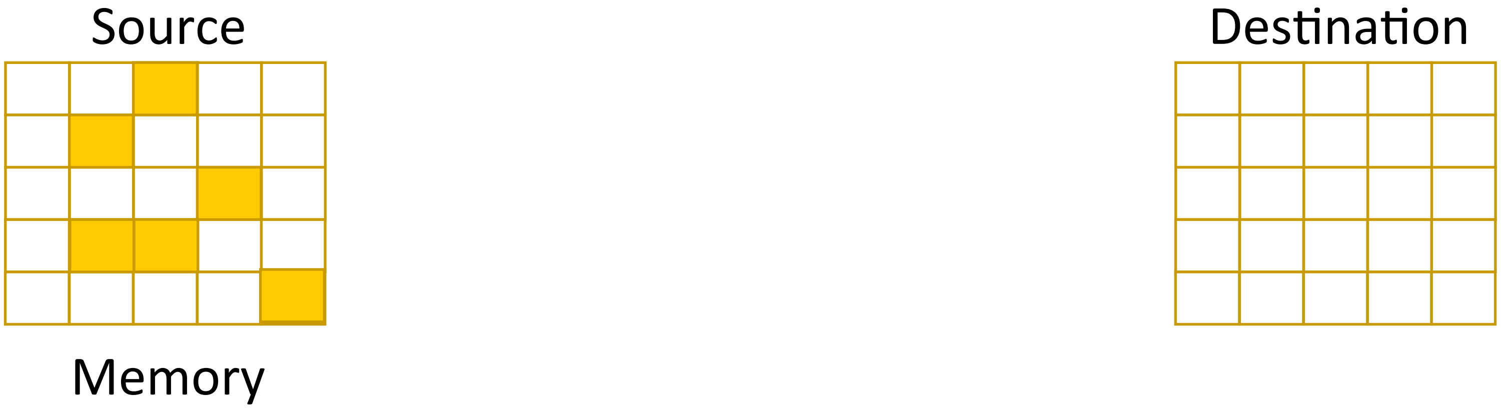
Live Migration:
Total Time vs Downtime

Source

Destination

Memory

Live Migration:
Total Time vs Downtime

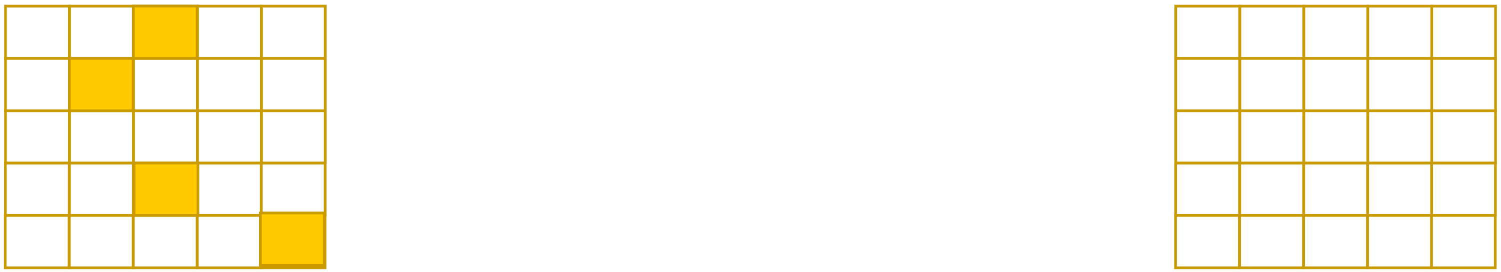


Source

Memory

Destination

Live Migration:
Total Time vs Downtime



Source

Destination

Memory

Live Migration:
Total Time vs Downtime

Source
Destination
Memory

- Iterative copy hopefully reduces the working set each iteration
- Depends on Network being faster than Dirty rate

Current Solution: Early Resume/Post-copy

- Problem: Applications can change memory faster than network bandwidth

- Solution:
  - Stop migration copy everything
  - Resume early if downtime is high

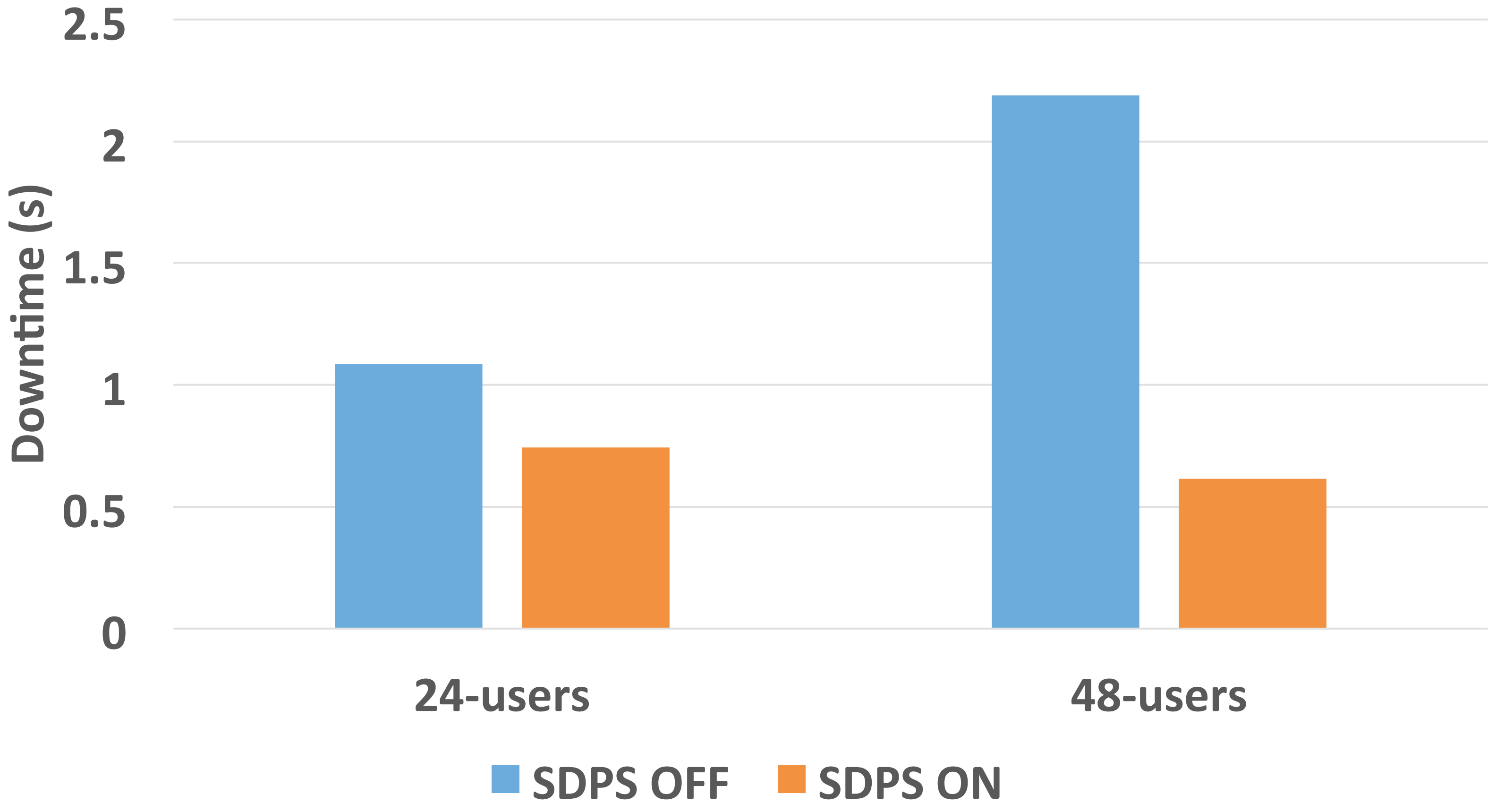
- Destination remote page faults on missing pages

Stun During Page Send (SDPS)

- Problems with Early Resume:
  - Remote page faults very expensive for high latency networks
  - Not atomic: Unsafe for WANs where network hiccups are more likely

- Solution:
  Throttle VM just enough to keep dirty rate < network rate

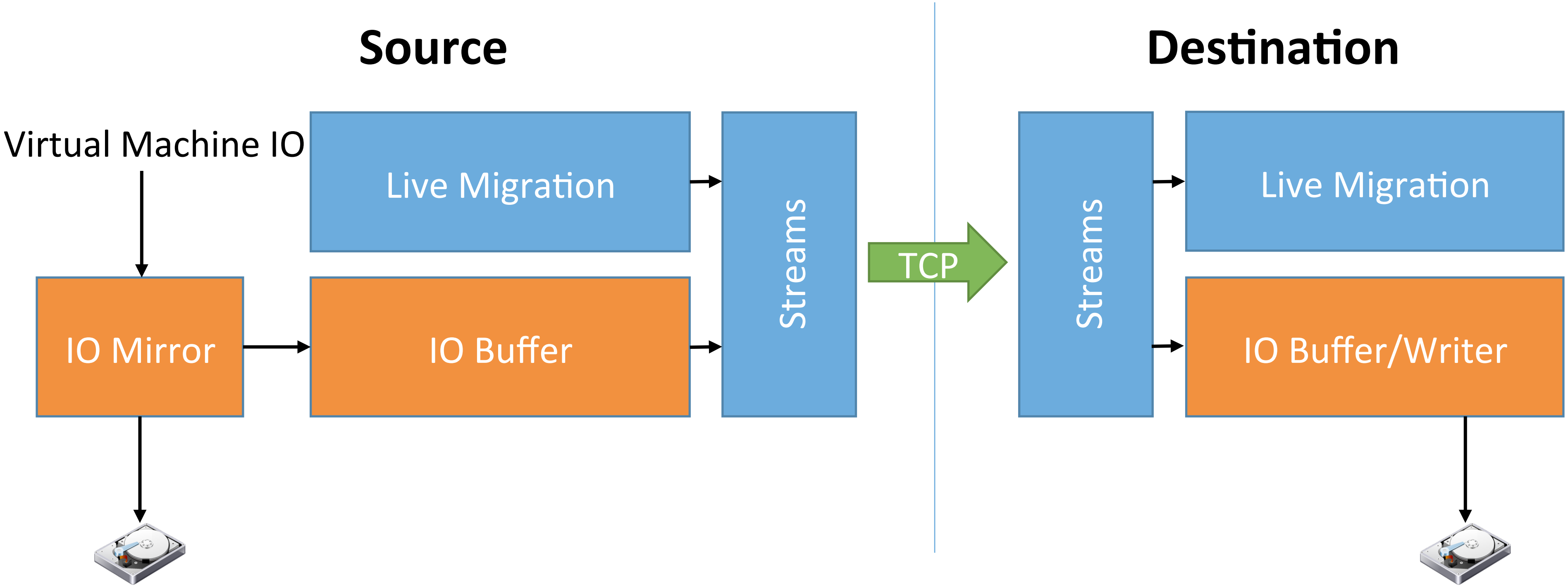
SDPS Downtime Results



Overview

- Architecture Overview
- Wide Area Memory Migration
- **Wide Area Storage Migration**
- Evaluation

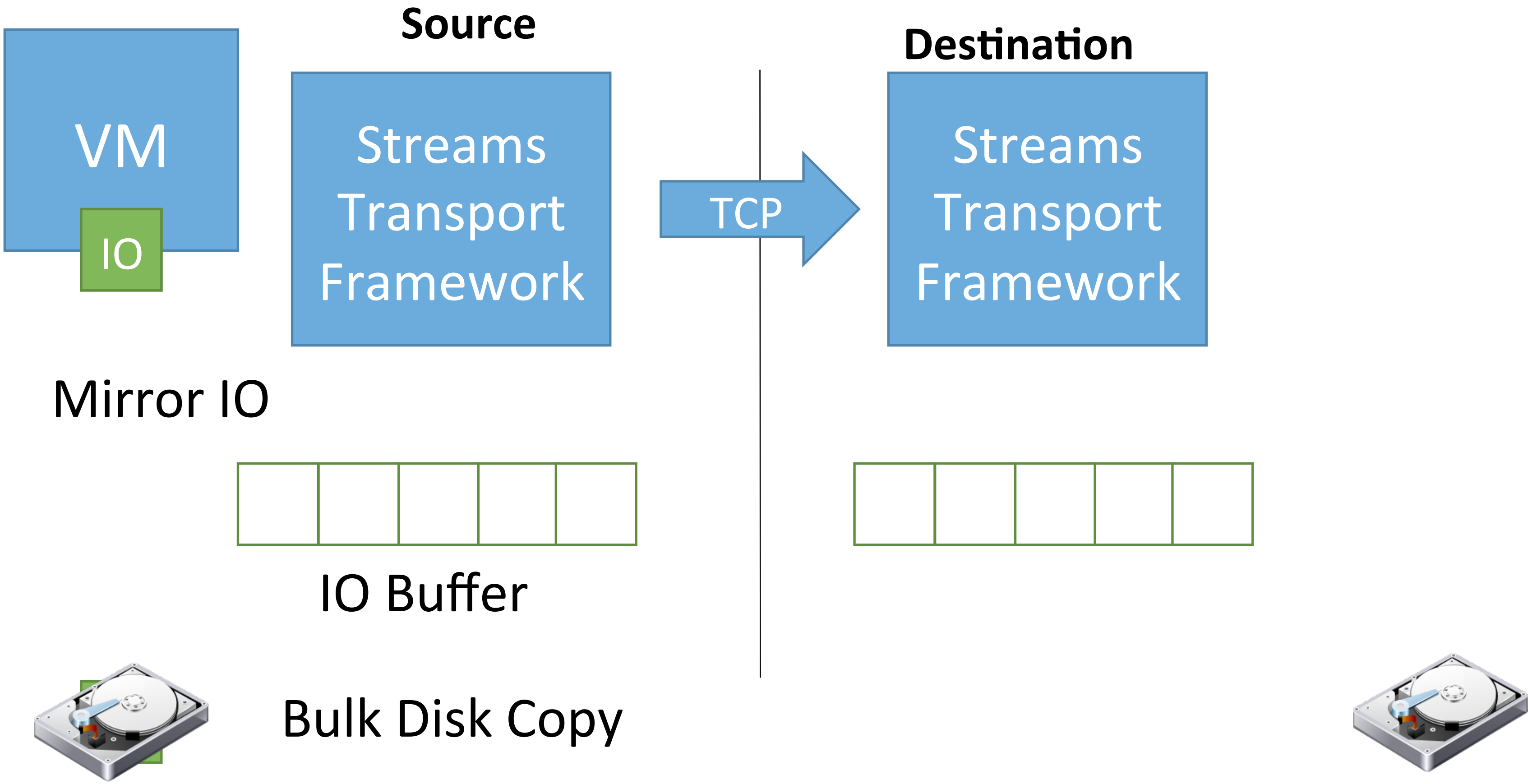
XvMotion Architecture: IO Mirroring



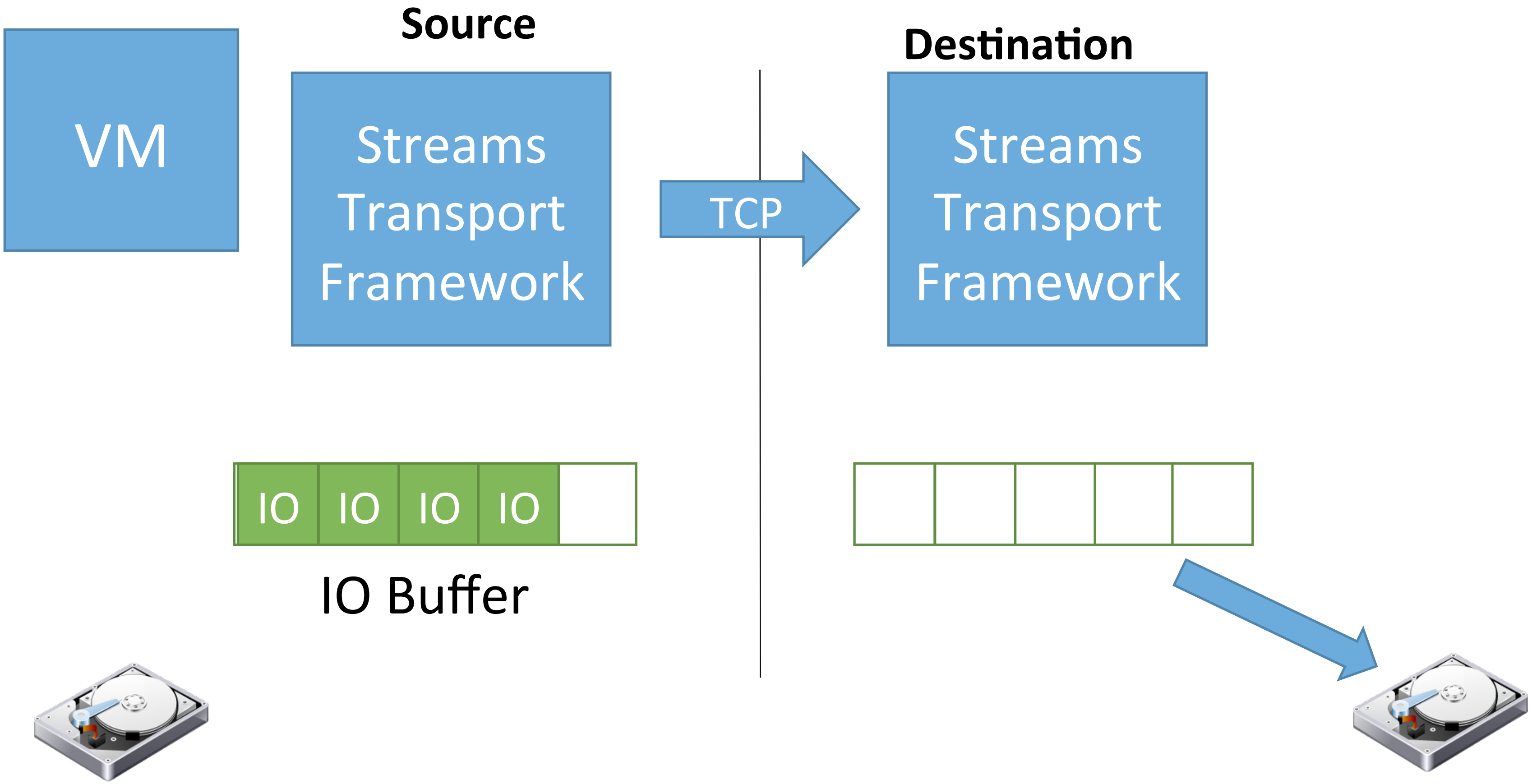
Problem: *Synchronous mirroring over the WAN slows guest workload*

Goal: *Hide network latency from VM*

XvMotion Buffering: Asynchronous IO



XvMotion Buffering: Asynchronous IO



Source

Destination

VM

Streams Transport Framework

TCP

Streams Transport Framework

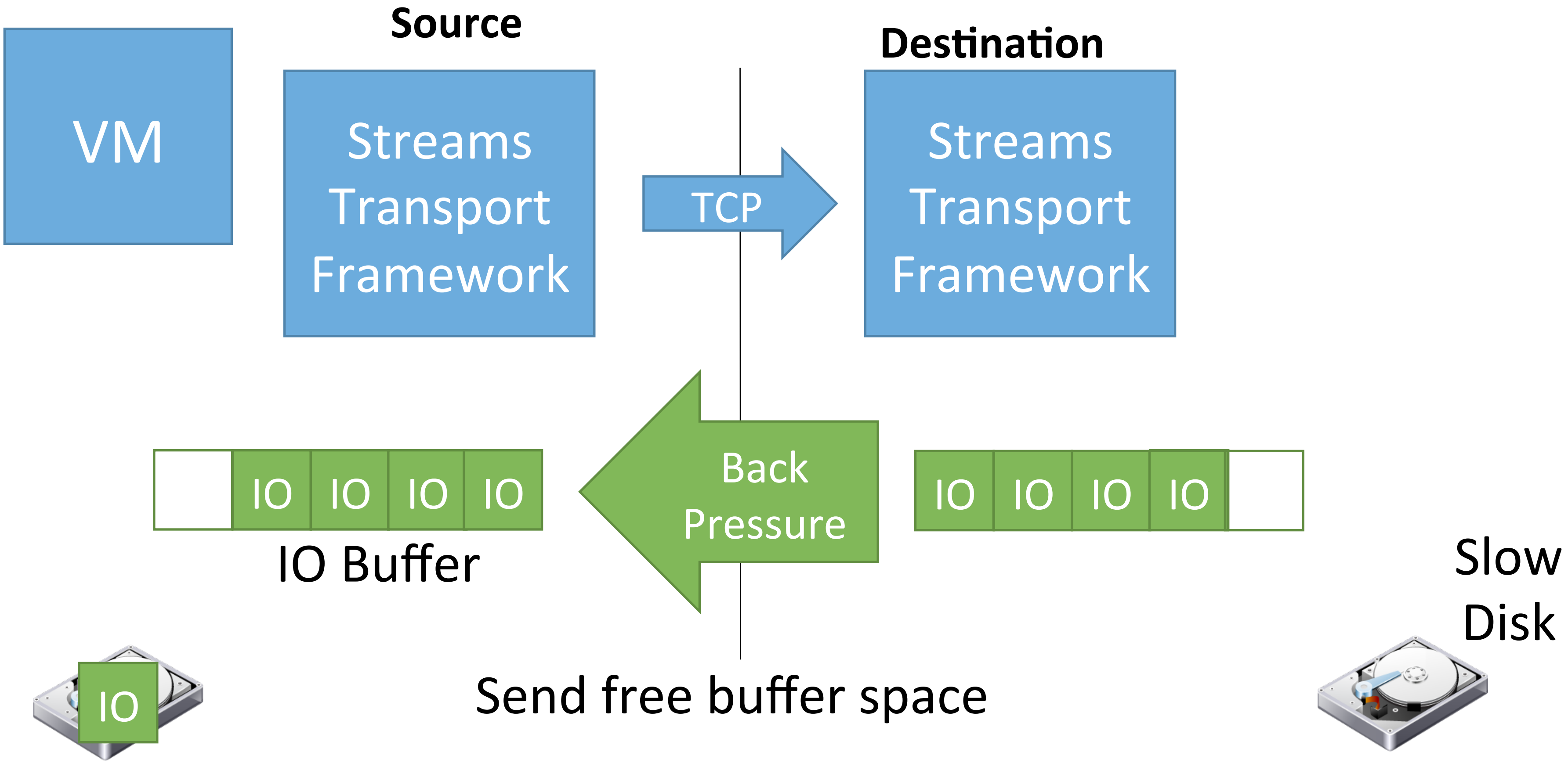
IO IO IO IO

IO Buffer

Problem: *Workload may be too fast on source for the destination*

Goal: Throttle copy process and workload as necessary

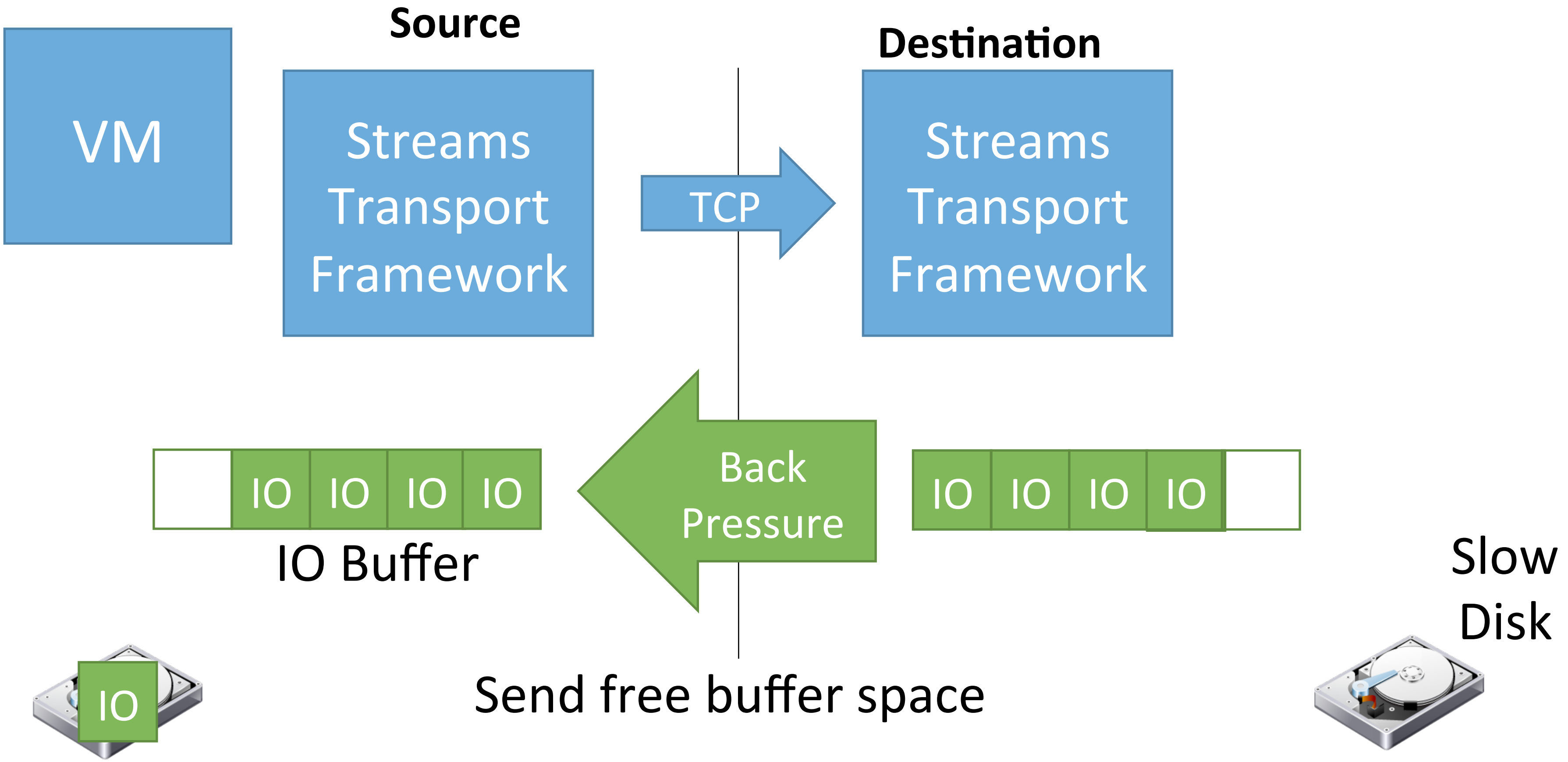
XvMotion Buffering: Congestion Control



Send free buffer space

Limit per-disk OIOs/buffer on destination

XvMotion Buffering: Congestion Control



Send free buffer space

Limit per-disk OIOs/buffer on destination

Overview

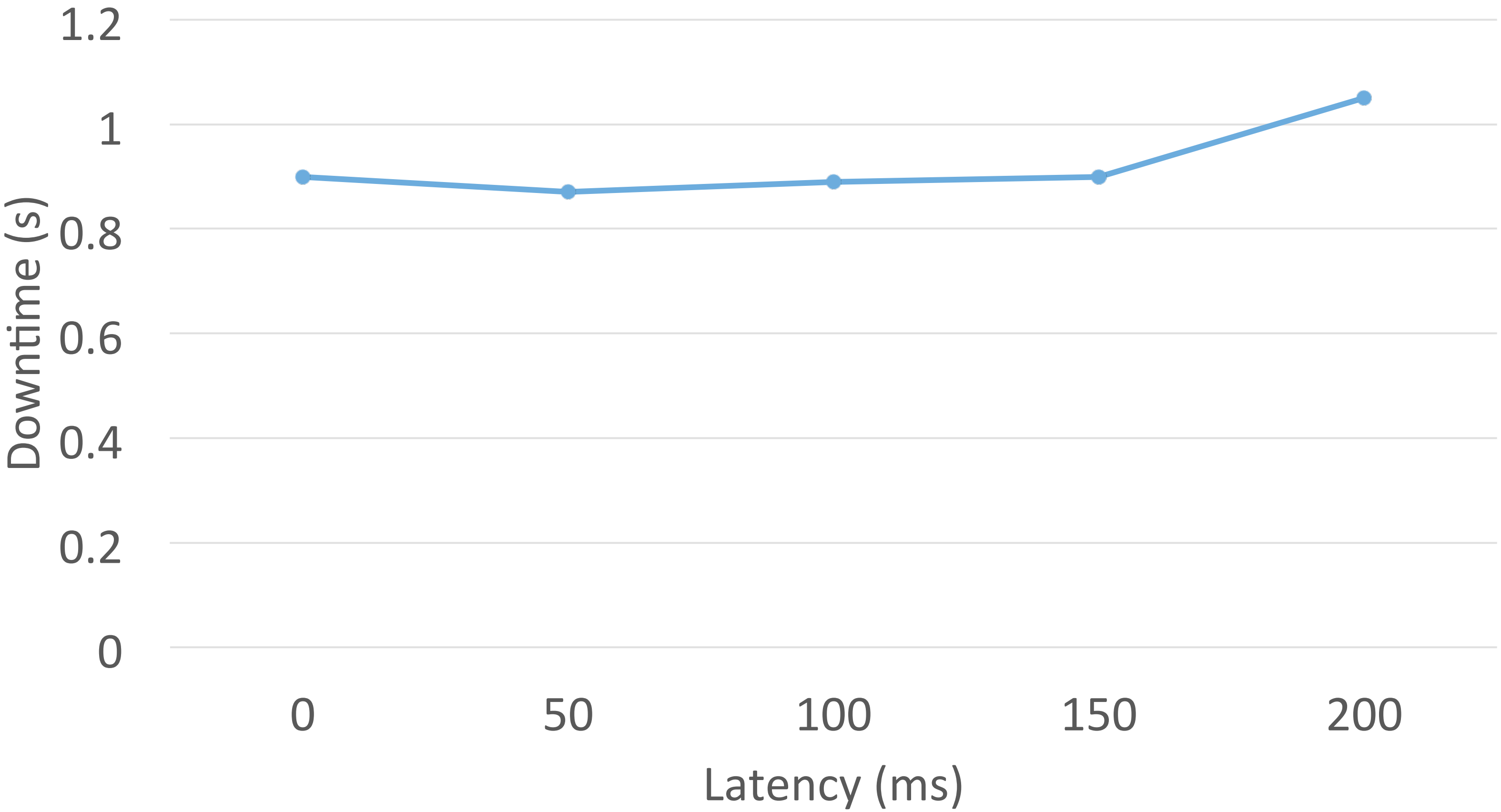
- Architecture Overview
- Wide Area Memory Migration
- Wide Area Storage Migration
- **Evaluation**

Evaluation

- Migration Time: Total end-to-end time
- Downtime: Time machine execution is suspended for final switchover
- Workload Penalty: Average percentage penalty to the VM workload

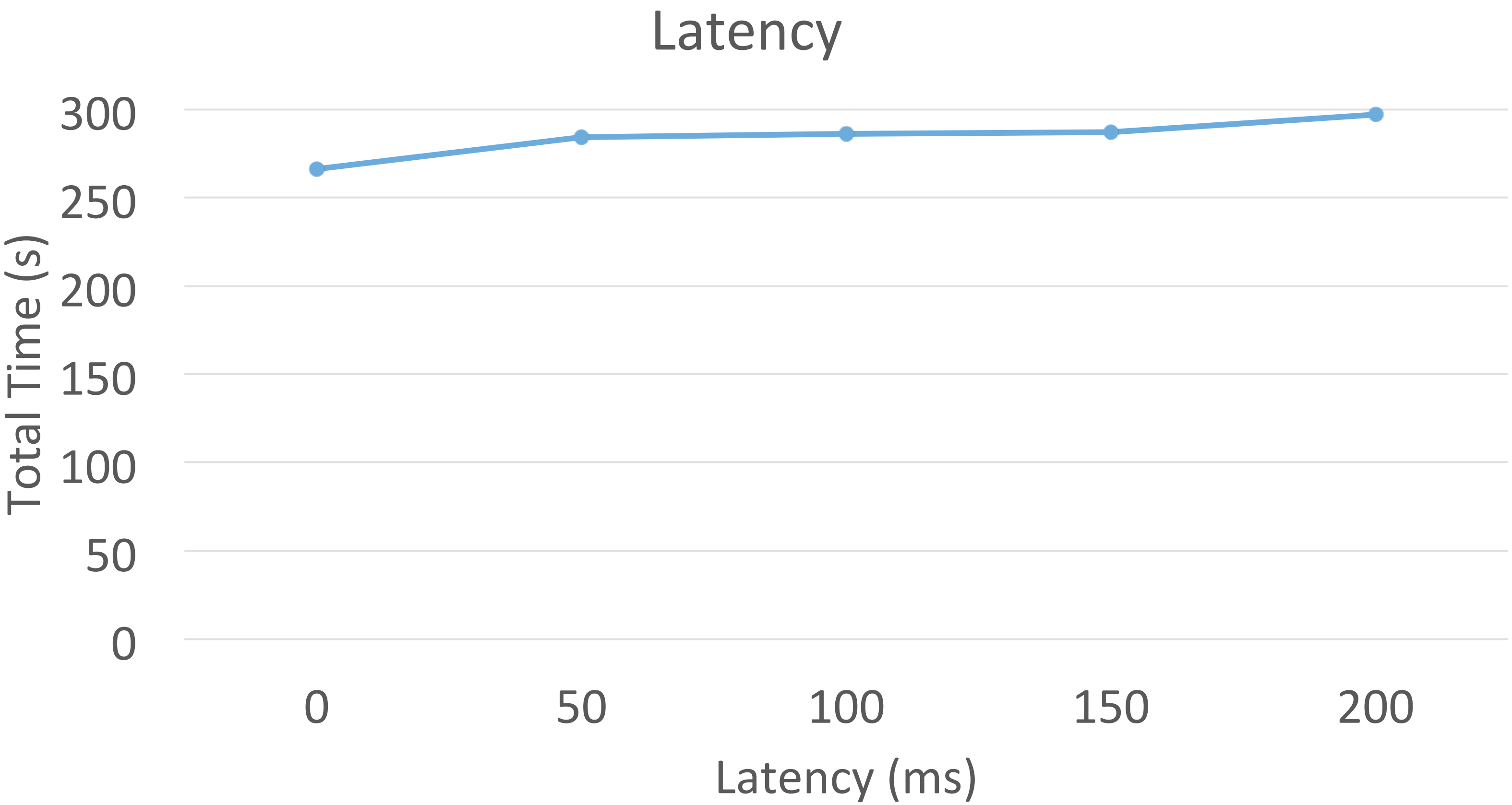
- OLTP IO Workload (data disk only)
- 10 GB system/12 GB data

XvMotion Downtimes



Take Away: ~1 second downtime independent of distance

XvMotion Migration Time

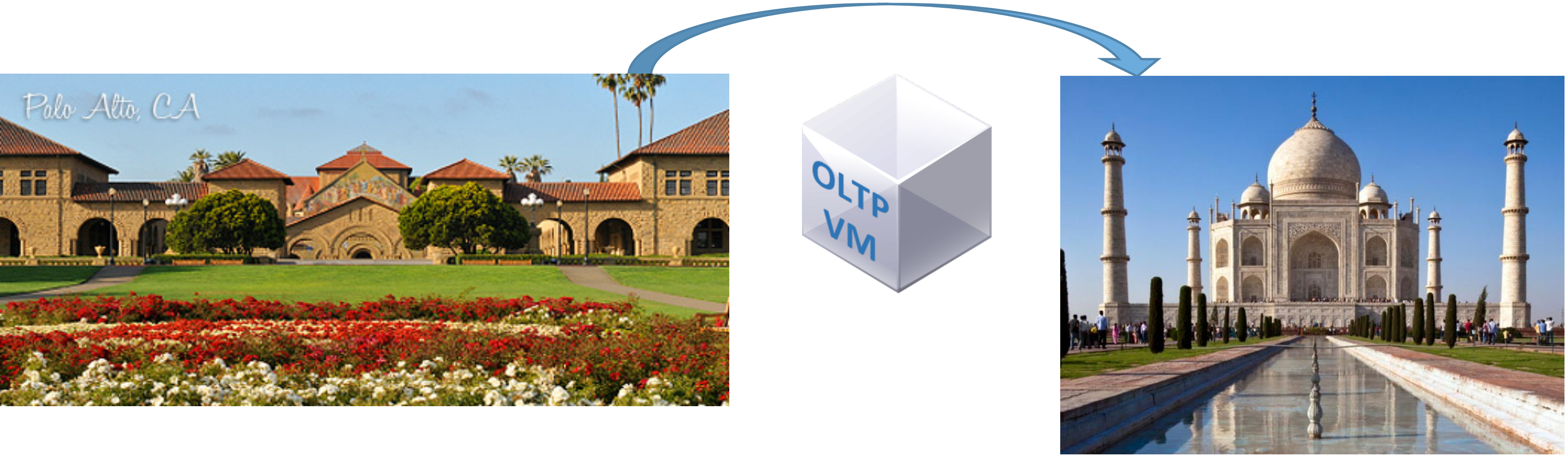


Take Away: small linear time increase with distance

California to India Migration

- 1 Gbps network with 214 ms RTT

- OLTP: 68 MB/s disk copy – 89 MB/s memory copy
- ~11% Workload Penalty from Throttling



Summary

- XvMotion frees migration from the need for shared storage and fast local networks
  - Tolerates wide area network bandwidth/latency and reliability
  - Tolerates heterogeneous storage performance
  - Downtimes and workload penalty comparable to local migration
  - Atomic Switchover
- Enables new use cases – e.g. disaster preparedness, cluster upgrade, shared nothing
- On the path to deployment:
  - Deployed in customer metro area networks
  - Cross continent migration e.g. Palo Alto to India is practical

Questions?