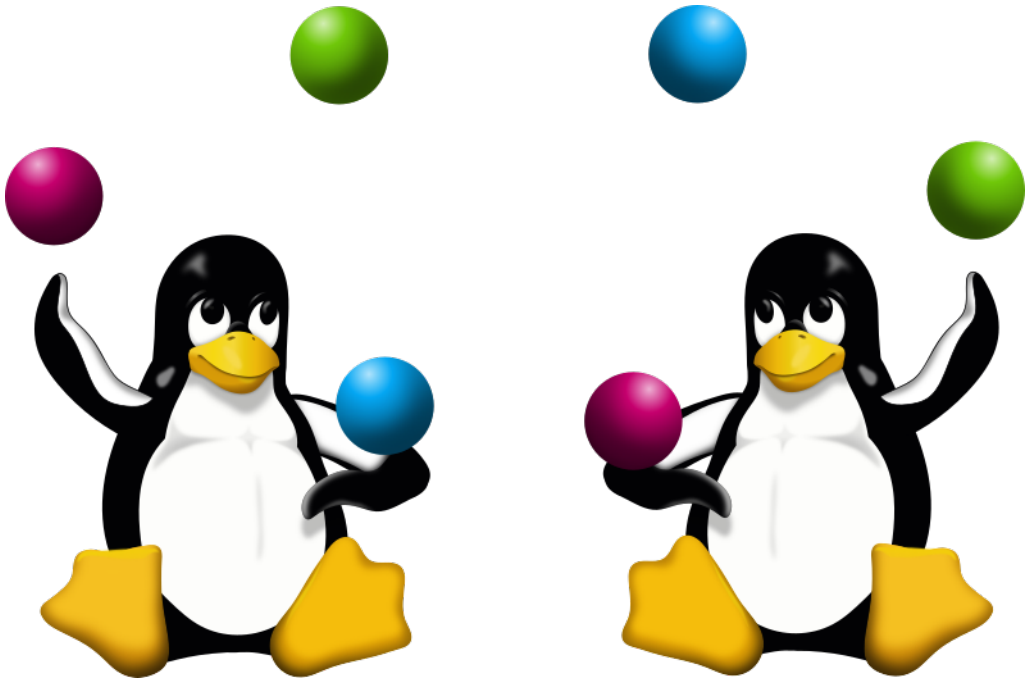
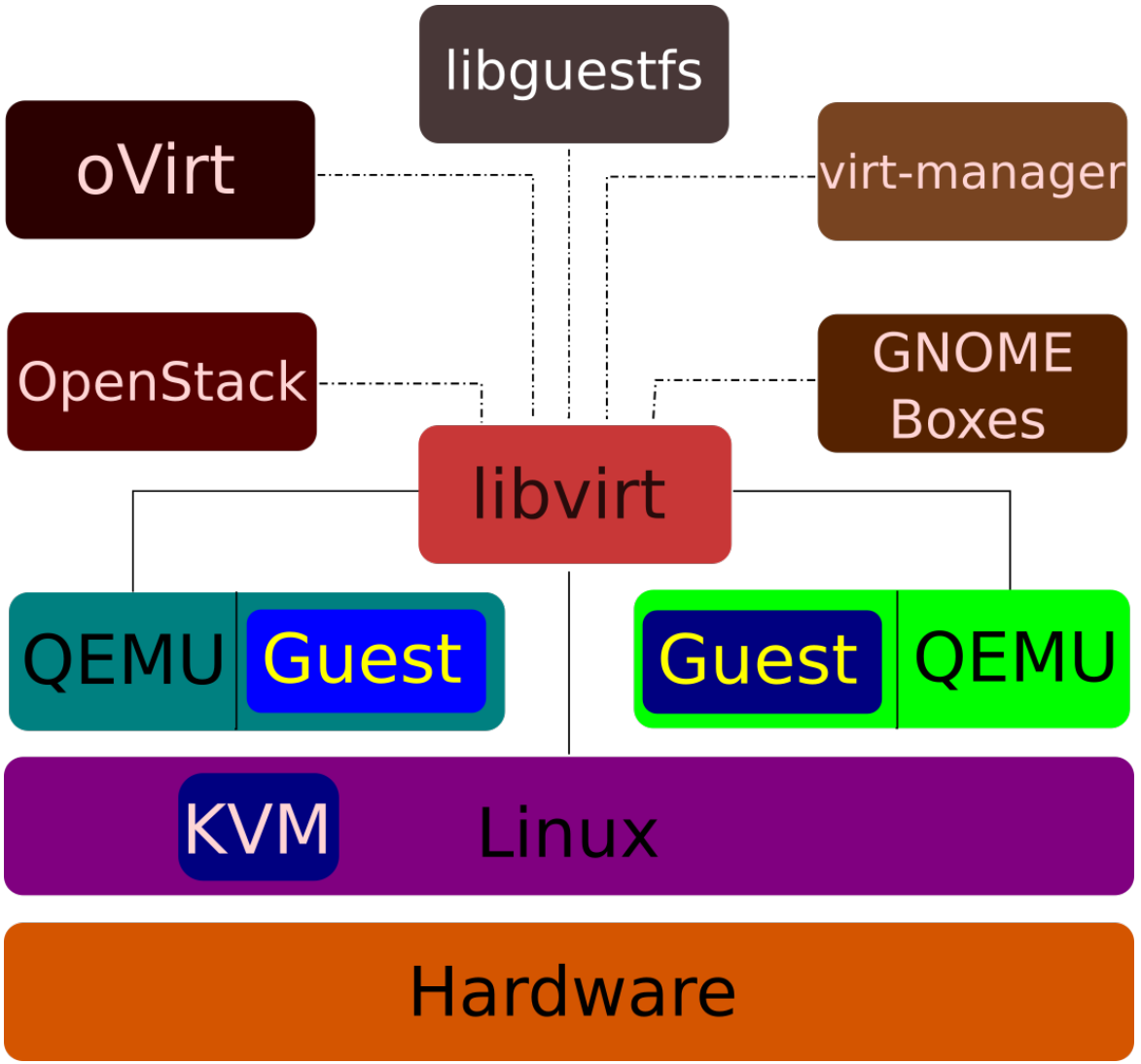
Live Migration of Virtual Machines From the Bottom Up



FOSDEM 2016

Amit Shah | Red Hat | amit.shah@redhat.com

Virtualization Stack



QEMU

- Creates the machine
- Emulates devices
  - some mimic real devices
  - some are special: paravirtualized
- Entire guest is contained within QEMU
- Uses several services from host kernel
  - KVM for guest control
  - Linux for resources
- Runs unprivileged

KVM

- Do one thing, do it right
- Linux kernel module
- Exposes hardware features for virtualization to userspace
- Emulates some devices
  - Like APIC
- Enables several features needed by QEMU
  - like keeping track of pages guest changes

libvirt

- Management of VMs, storage, network
- Provides a stable API
- Remote management
- `virsh` – command-line interface
- `cgroups`
- `sVirt`
- Makes it possible for QEMU to run unprivileged
  - Opens files, connections and passes them on to QEMU

Note on higher layers

- OpenStack
  - Cloud or IaaS management
- oVirt
  - Data centre management
- virt-manager / GNOME Boxes
  - PC management
- libguestfs
  - nifty tool to perform several operations on VM images

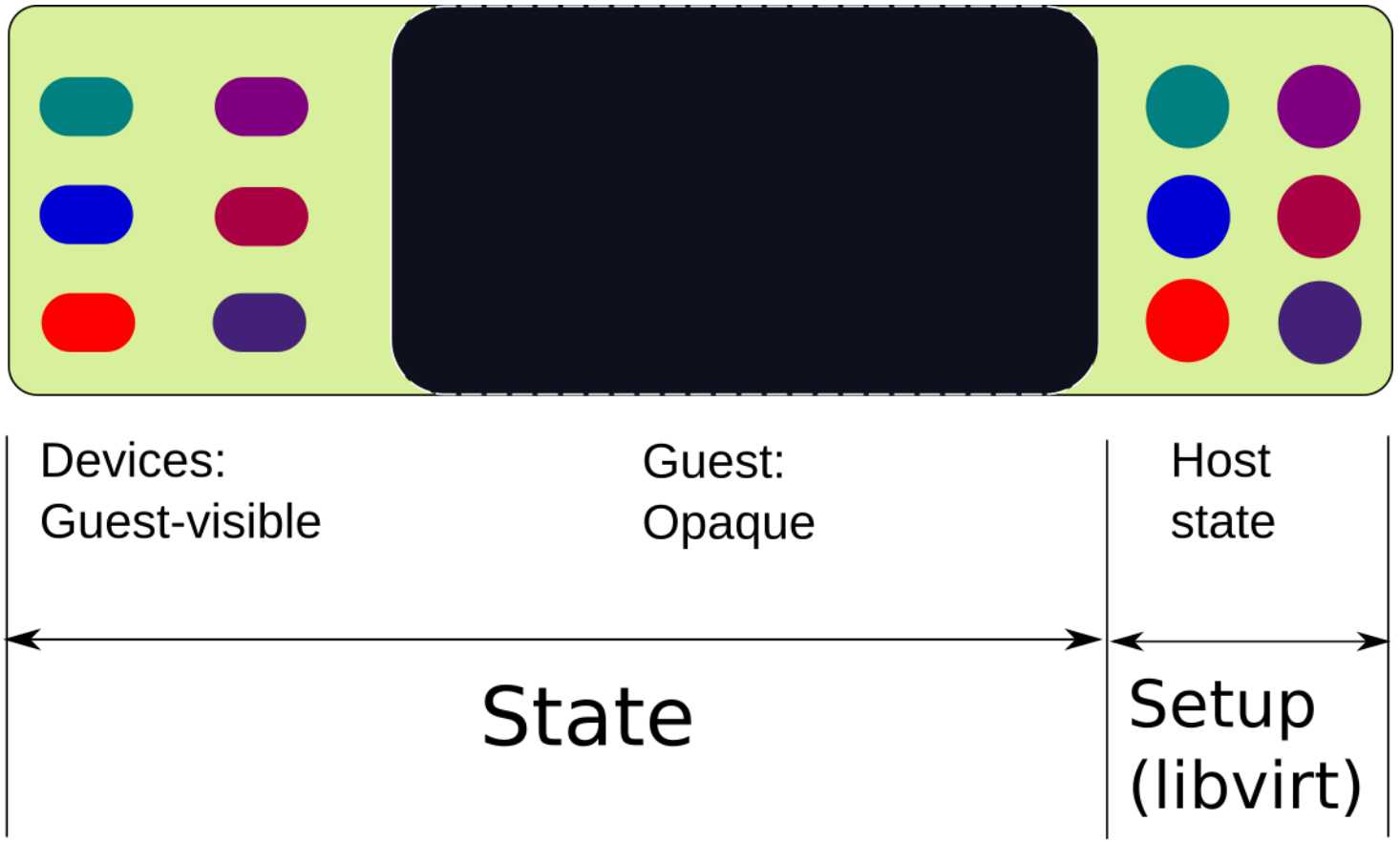
KVM Today

- Very good performance and scalability
  - Consistently tops SPECVirt results
- Default hypervisor for oVirt, OpenStack
- Out-of-box support in all distributions

Live Migration

- Pick guest state from one QEMU process and transfer it to another
  - while the guest is running
- The guest shouldn't realize the world is changing beneath its feet
  - in other words, the guest isn't involved in the process
  - might notice degraded performance, though
- Useful for load balancing, hardware / software maintenance, power saving, checkpointing, ...

QEMU Layout


Devices:
Guest-visible
Guest:
Opaque
Host
state
State
Setup
(libvirt)

KVM Today

- Very good performance and scalability
  - Consistently tops SPECVirt results
- Default hypervisor for oVirt, OpenStack
- Out-of-box support in all distributions

Workstations

Generic-office-desktop by averpix, https://openclipart.org/detail/127213/genericofficedesktophttps://openclipart.org/detail/127213/genericofficedesktop

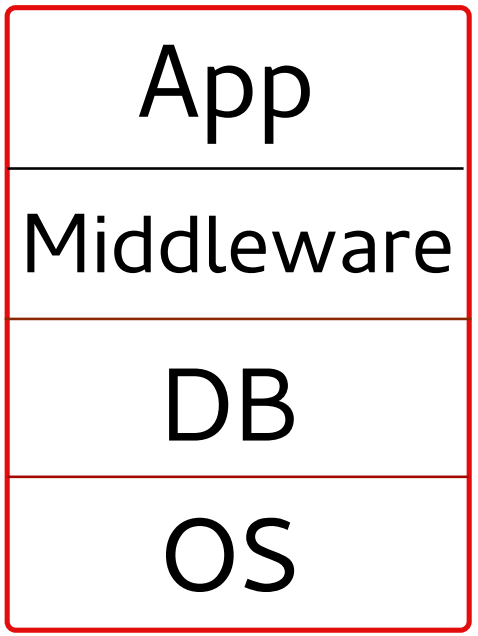
- Main interaction with guests
- Migration is triggered by admins
- Don't need anything more fancy / heavyweight

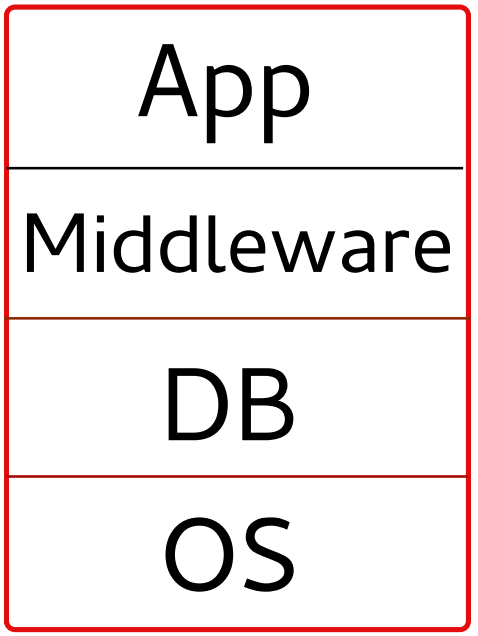
Data Centres / Clouds

Server-1U by Rob Fenwitch, https://openclipart.org/detail/169833/server-1uhttps://openclipart.org/detail/169833/server-1u

- Main interaction with hosts
- Migration is triggered by policies, transparent to admins
- Policies optimise resource usage; host maintenance, etc.

Data Centres

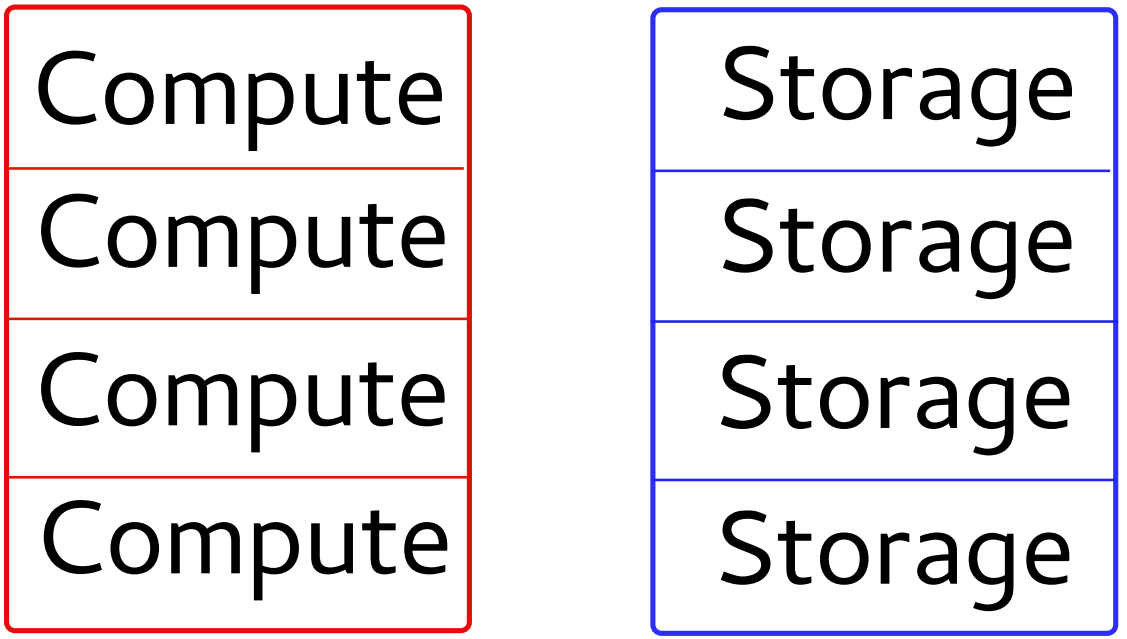




- Scale-up
- Traditional workloads
  - large databases

- Many vCPUs
- Lots of RAM
- Critical data
- Shared storage

IaaS / Clouds



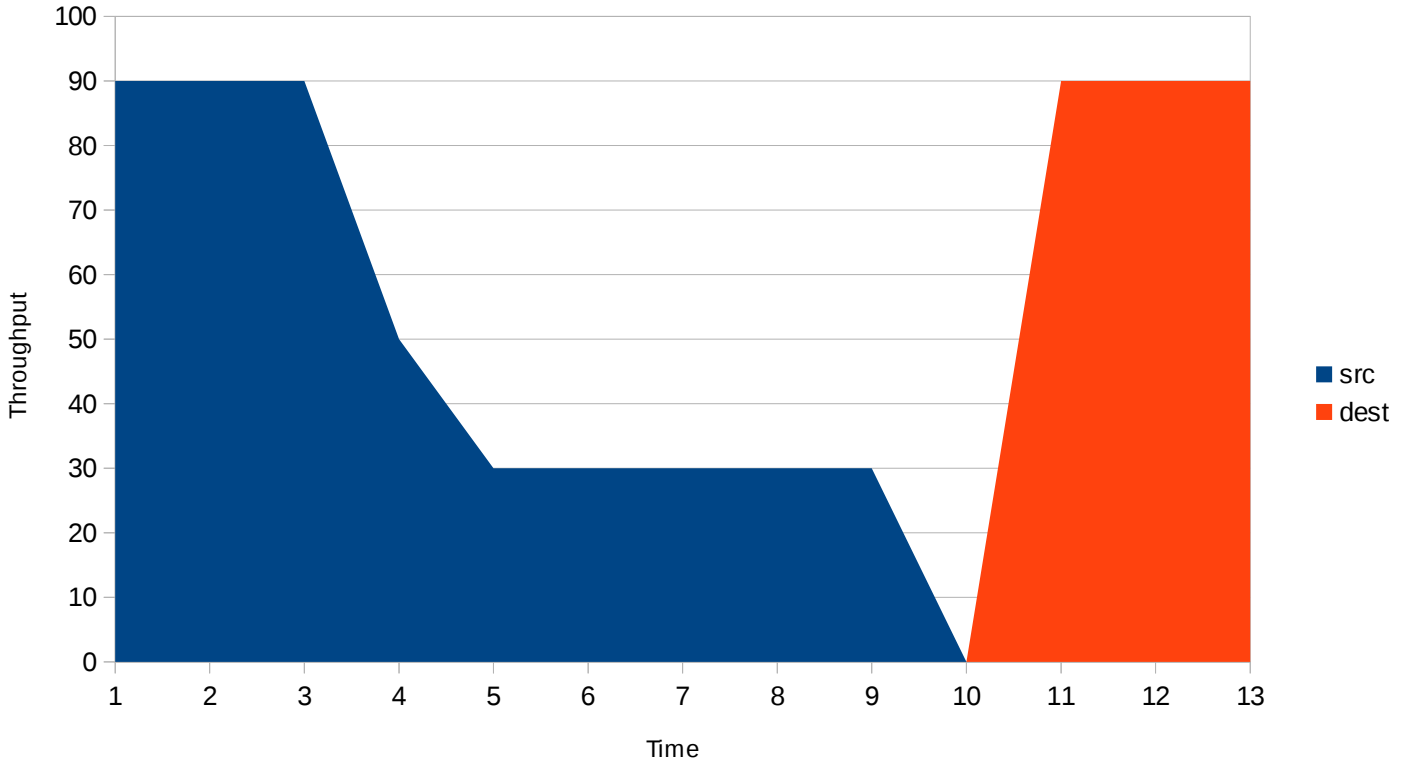
- Scale-out
- Custom applications

- Compute, storage separate
- Sometimes compute has storage
  - Needs block migration

Block Migration

- Using only QEMU
  - Take a snapshot of disk image
  - Migrate base disk image
  - Migrate RAM and new snapshot
  - Iterate till VM converges
- Using libvirt
  - Setup NBD connection between hosts
  - Transfer block device contents across hosts

Big VMs



- Performance drop while migration in progress
- Customers don't like this

QEMU Main Loop (old)

```
main_loop()
{
  while (1) {
    service_guest_requests();
    service_guest_io();
    migration_pass();
  }
}
```

QEMU Main Loop (new)

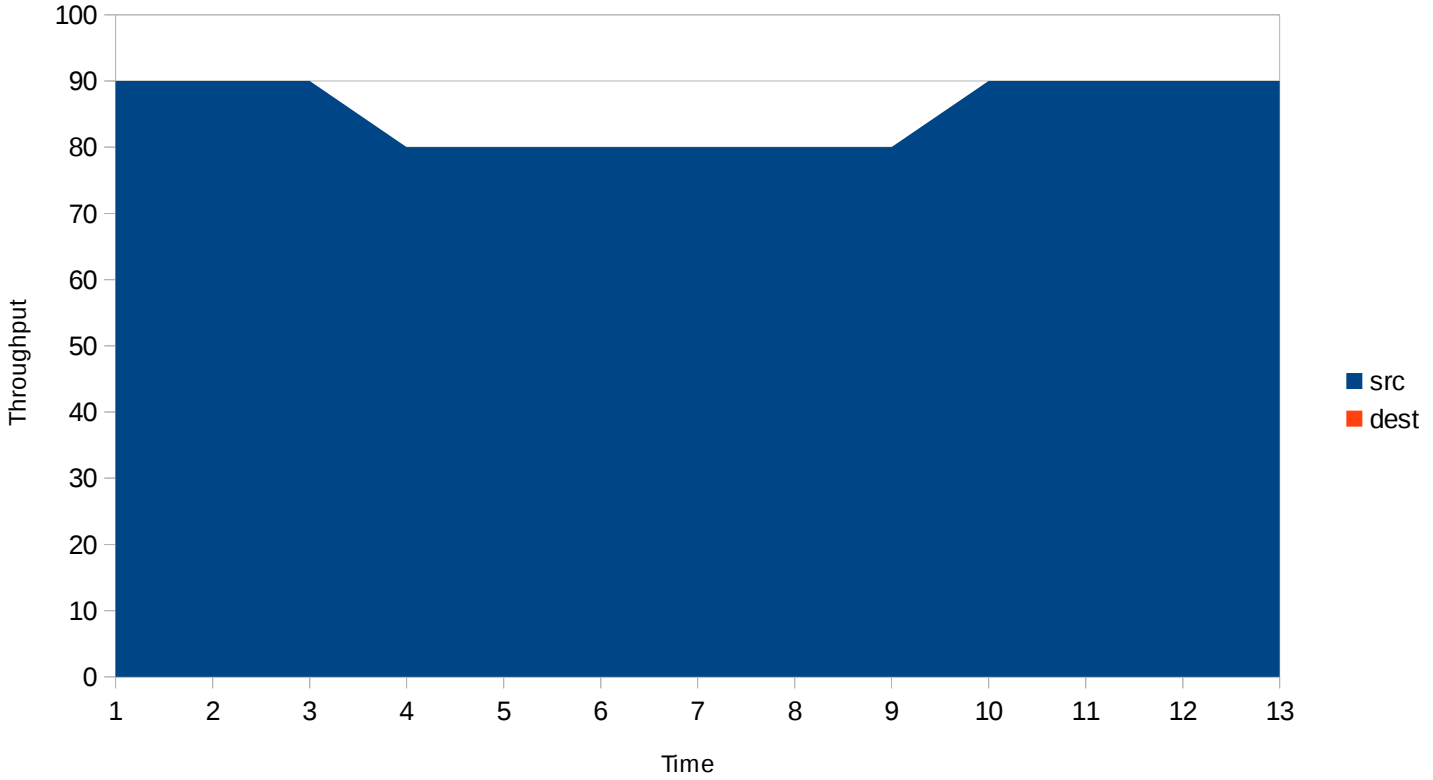
thread1

```
main_loop()
{
    while (1) {
        service_guest_requests();
        service_guest_io();
        }
}
```

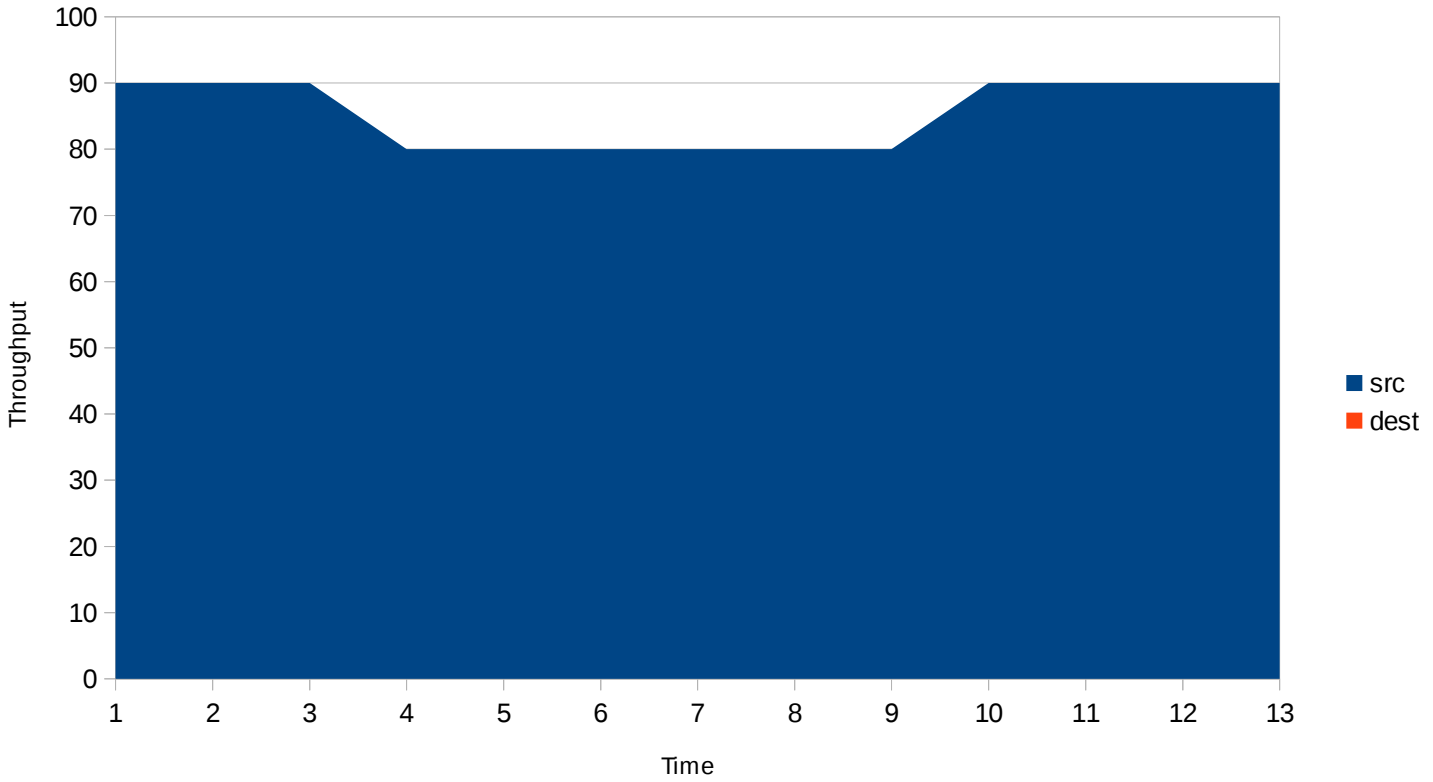
thread 2

```
migration_thread()
{
    while (1) {
        migration_pass();
    }
}
```

- Add migration thread
- Helps parallelise guest IO and migration passes


100
90
80
70
60
50
40
30
20
10
0
Throughput
1
2
3
4
5
6
7
8
9
10
11
12
13
Time
src
dest

Oops



- Guest doesn't migrate
- Slowness was a feature!

Restrict Guest

- Throttle guest vCPUs
  - Hope the rate of dirtying memory reduces
  - Autoconverge
  - cgroups
- Offline guest vCPUs

Compression

- Multi-threaded compression
  - Compress pages before sending
  - Do this in multiple threads
- xbzrle
  - Send diffs of pages from previous iteration
  - Means we have to maintain a cache of pages sent in previous iteration

Postcopy

- Migrate guest before all RAM has been transferred
- Keep transferring pages from src to dest on a new channel
- Remote-page-fault pages which don't exist on dest
  - Special OOB mode of transferring pages on the new channel
- userfaultfd in Linux implements remote page fault functionality

Other Challenges

- Multiple migrations
  - Logs get left behind on older hosts
  - 24th migration might be failing, 23 prior ones have succeeded
    - but we don't know it's the 24th attempt
- Multiple layers
  - Have to check logs for each layer top->down to find cause
- QEMU defaults
  - Not suitable for all projects
  - QEMU devels don't know about deployment scenarios
  - More communication between projects to understand options
  - New focus on feature pages to expose more info to higher levels

Thank You!

Amit Shah | http://log.amitshah.net | amit.shah@redhat.com