Live Migration
- Allows serving customers while migrating
- When done in-band uses the same bandwidth as users’ request
- Therefore, we expect degradation in the quality of service being migrated
- Focus on migrating a single service from a source host to a destination host and would like to minimize the cost of live migration

Migration Algorithm
- Live migration has two phases:
  - pre-copy phase where pages are transferred while the service is alive at the source server
  - stop-and-copy phase where the service is stopped and the rest of the pages are copied
- Pages that were copied in the pre-copy phase may get dirty later on and need to be transferred again
- The migration algorithm needs to determine:
  - How much bandwidth to use during the pre-copy phase
  - When to move to the stop-and-copy phase
- The following graph depicts the total cost of a live migration using different (fixed throughout the migration process) bandwidths
- There is a non-trivial trade-off between minimizing the copy phase duration and maintaining an acceptable quality of service during the pre-copy phase

The Cost of Live Migration
- Previously evaluated solely on the downtime
- We also consider the migration impact on the service performance in terms of SLAs
- To study this affect we studied service quality in terms of SLA when the available bandwidth is reduced
- We model a web server as an M/M/1 queue, then the cost function can be calculated using the following formula:
  \[ e^{t(\lambda-\mu(B_s))} \] (\( t \) is a response time enforced by the SLA, \( \lambda \) is the request rate and \( \mu \) is the service rate)
- This graph depicts the degradation of a running web server (Tomcat) as we limit the available bandwidth

Results
- We developed two algorithms:
  - CALM-fixed, when the bandwidth used for migration is fixed throughout the migration process
  - CALM-adaptive, when the bandwidth is adjustable during the migration process
- Compared with the XEN’s migration algorithm
- Two different environments:
  - Service consumes small fraction of the available bandwidth
  - Service consumes most of the available bandwidth