Live Migration of Virtual Machines

Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen†,
Eric Jul†, Christian Limpach, Ian Pratt, Andrew Warfield

NSDI 2005

Outline

- Introduction
- Related Work
- Design
- Writable Working Sets
- Implementation
  - Managed Migration
  - Self Migration
  - Dynamic Rate-Limiting
  - Some implementation issues
- Evaluation
- Future work
- Conclusion

Introduction [1/2]

- The research field of operating system virtualization is receiving much attention in data center and cluster computing
  - paravirtualization
- Migration of entire virtual machine
  - Original host are terminated after migration
    - Maintenance and repair of original host
  - In-memory state can be maintained consistently
    - No restart and reconnection
  - Migration concerns between users and operators is simplified

Introduction [2/2]

- Issues
  - Downtime and total migration time
  - Resource contention
    - CPU
    - Network bandwidth
- Approach
  - Pre-copy
    - Migration without stop
  - Stop-and-copy
    - Migration when the VM is not running

Related Work

- Collective project
  - Migration of an OS instance for mobility
  - Different physical hosts and different times
- Zap
  - Migration of process domains using partial OS virtualization
  - No live migration
- NomadBIOS
  - Pre-copy migration
  - No rate adaptation of the writable working set
- Process migration
  - Residual dependencies
- Sprite
- MOSIX

Design [1/3]

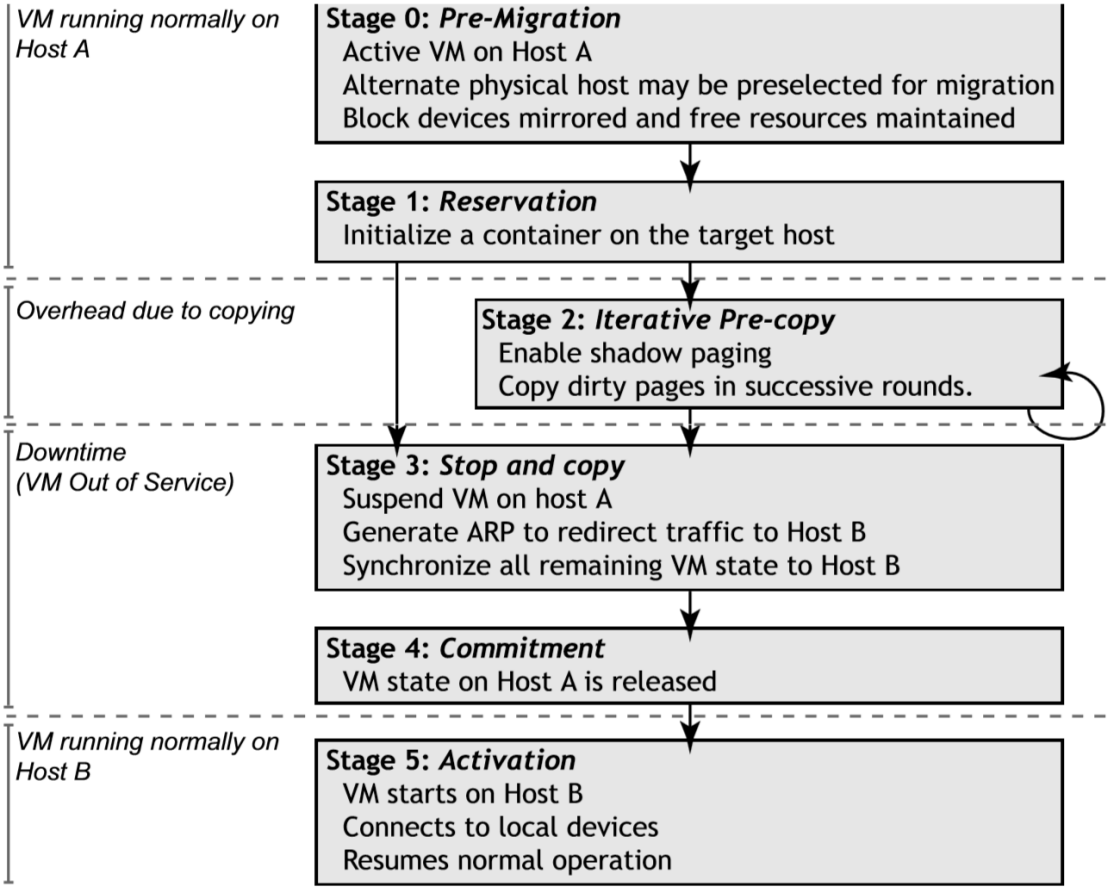
- Migrating Memory
  - Balance of downtime and total migration time
    - Downtime: time when service is not available
    - Total migration time: all time during migration
  - Bounded Iterative push phase
    - pre-copy
  - A short stop-and-copy phase
  - Writable working set (WWS)
    - Adjust the page ratio of pre-copy phase and stop-and-copy phase
    - WWS behavior
  - Service degradation can be prevented by adjusting network and CPU resources appropriately

Design [2/3]

- Network resources
  - Local area network
  - Advertise IP movement using unsolicited ARP reply
    - In some environments, the operating system will reply directly after recognizing the migration
- Local storage
  - Network-attached storage (NAS)
    - NAS allows all hosts to access storage
  - Local disk storage

Design [3/3]

- Migration timeline
  - OS migration from host A to host B

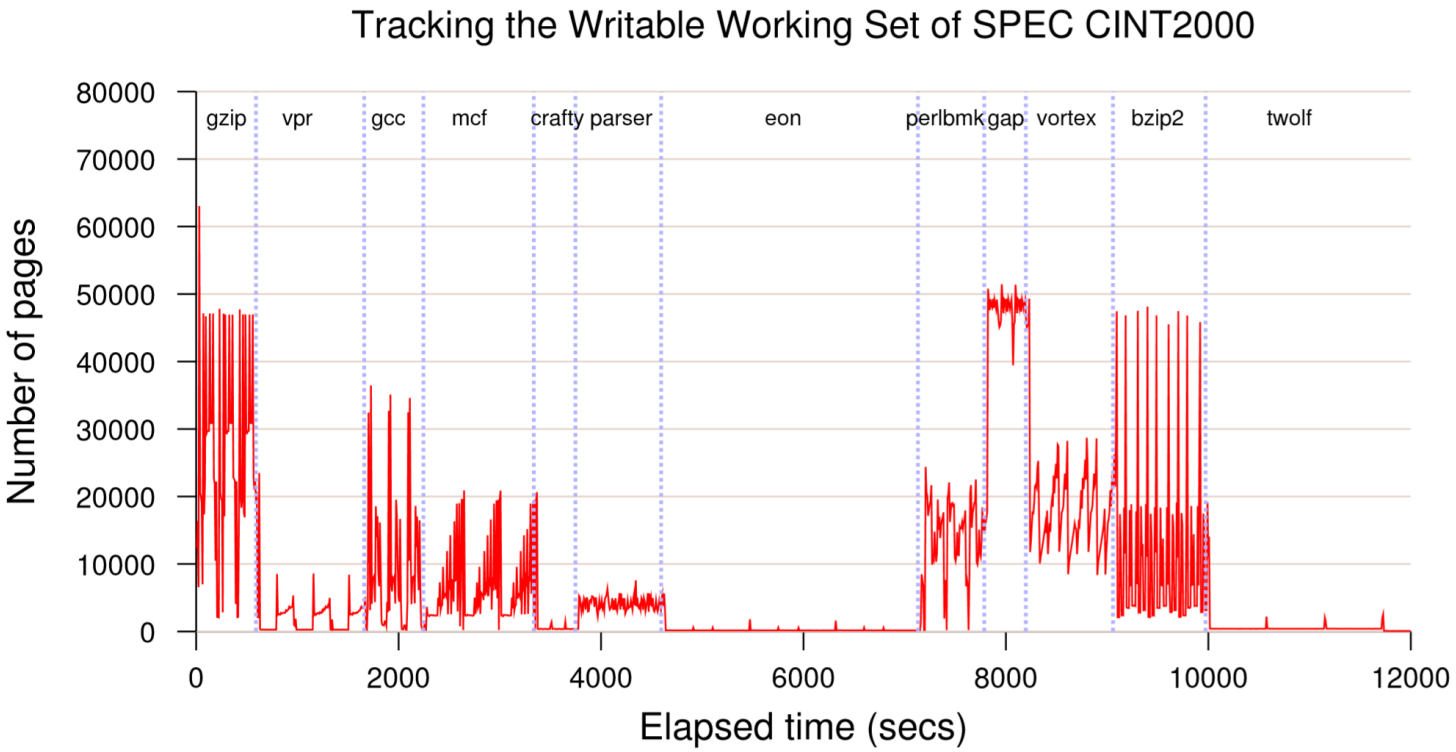


Writable Working Sets [1/2]

- Frequent modification of memory pages causes the overhead of page transfer
- Methods for determining the stop time for pre-copy approach is needed
  - No page modification
    - 1 pre-copy
  - Frequent page modifications
    - All pre-copy are vain
    - Stop and copy is required
- Writable Working Sets
  - Determine the page set for pre-copy approach and the page set for stop-and-copy (WWS)

Writable Working Sets [2/2]

- Measuring Writable Working Sets
  - A serious of programs that run over a period of time
  - Page size: 4KB



Managed Migration

- Performed by migration daemons running in the management VMs of the source and destination hosts
- Use shadow page table to track dirty pages in each push round
  - Xen inserts shadow page table under the guest OS
  - All PTEs are initially marked read-only
  - If the guest tries to modify a page, the resulting page fault is trapped by Xen
  - Xen checks the OS's original page table and forwards the appropriate write permission, marks the page as dirty in the bitmap
- At the beginning of next push round
  - The bitmap is copied to the control software, Xen's bitmap is cleared
  - The shadow page tables are destroyed and recreated, all write permissions are lost

Self Migration

- Implemented almost entirely within the migratee OS with only a small stub required on the destination machine
- No modifications are required either to Xen or to the management software running on the source machine

Difference	Managed Migration	Self Migration
Track WWS	Shadow page table + Bitmap	Bitmap + A spare bit in PTE
Stop-and-copy	Suspend OS (to obtain a consistent checkpoint)	2-stage stop-and-copy Ignore page updates in last transfer

Dynamic Rate-Limiting

- Dynamically adapt the bandwidth limit during each pre-copying round
- The administrator selects a min and a max bandwidth limit
- The 1st pre-copy round transfers pages at the minimum bandwidth limit
- Each subsequent round
  - Dirtying rate + constant increment (50Mbits/sec)
  - Dirtying rate = $\frac{the\ \#\ of\ pages\ dirtied\ in\ the\ previous\ round}{duration\ of\ the\ previous\ round}$
- Pre-copying terminated when
  - Calculated rate > max bandwidth
  - Less than 256KB remains to be transferred
- During the final stop-and-copy phase, transfers memory at the maximum allowable rate

Some Implementation Issues

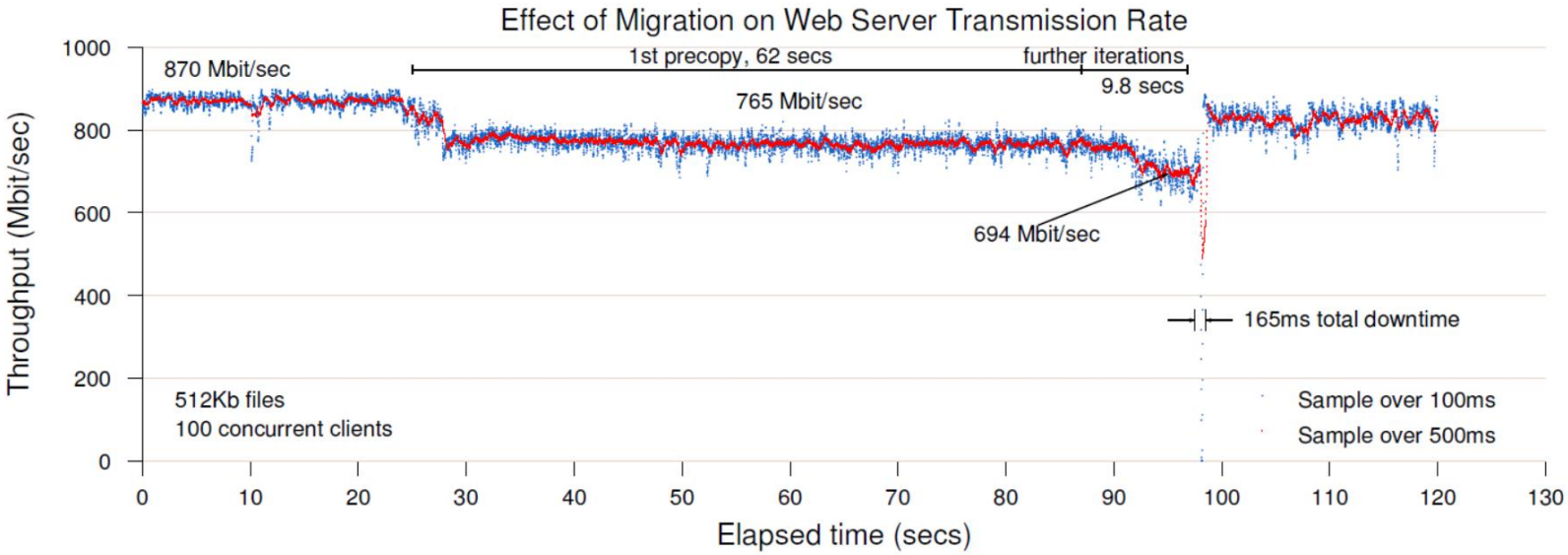
- Rapid Page Dirtying
  - Page dirtying is often physically clustered
  - Periodically ‘peek’ at the current round’s dirty bitmap
  - Transfer only those pages dirtied in the previous round that have not been dirtied again at the time we scan them
- Stunning Rogue Process
  - Fork a monitoring thread within the OS kernel when migration begins
  - If the process dirty memory too fast, then ‘stun’ it
- Freeing Page Cache Pages
  - OS can return some or all of the free pages
  - Do not transfer these pages while the 1st iteration

Evaluation

- Test Setup
  - Dell PE-2650 server-class machine
  - Dual Xeon 2Ghz CPUs
  - 2GB memory
  - Broadcom TG3 network interfaces
  - Gigabit Ethernet
  - Storage: iSCSI protocol from an NetAPP F840 NAS
  - XenLinux 2.4.27

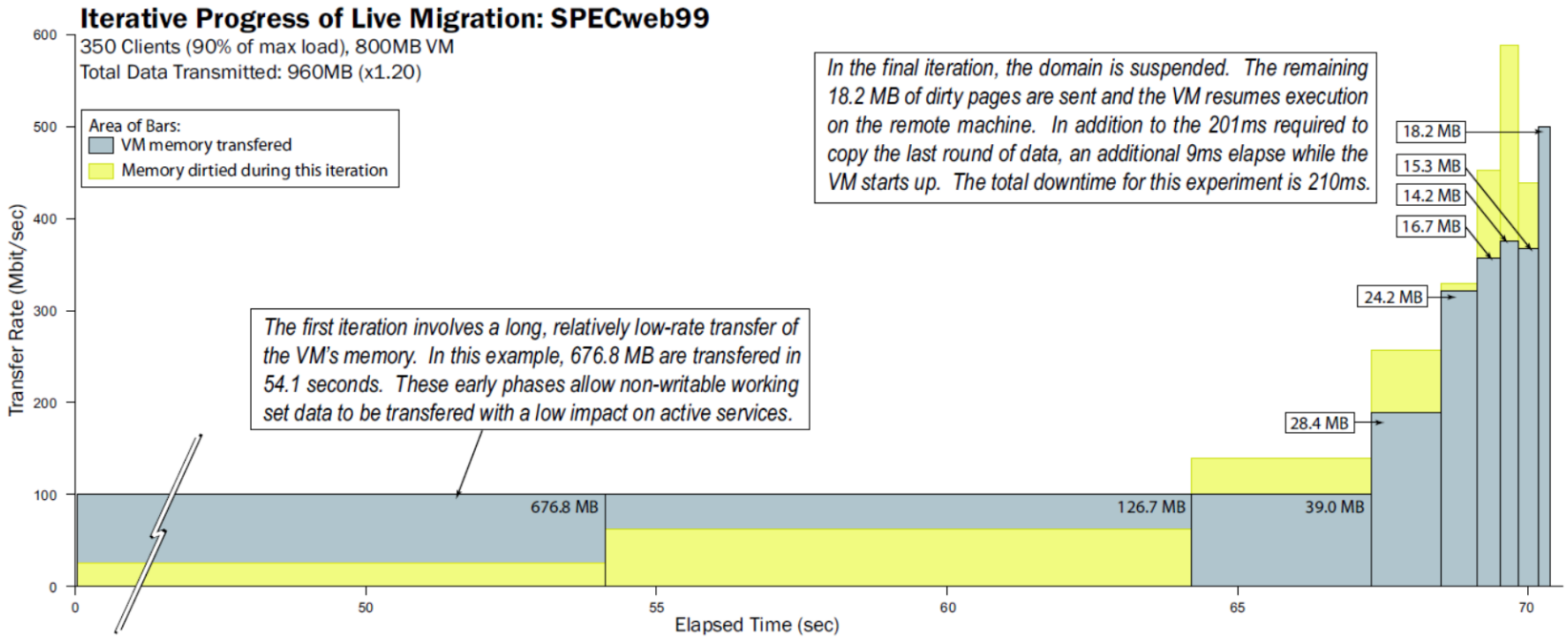
Apache 1.3 Web server

- Simple Web Server
- Continuously serving a single 512KB file to a set of 100 clients
- Total downtime = 165ms



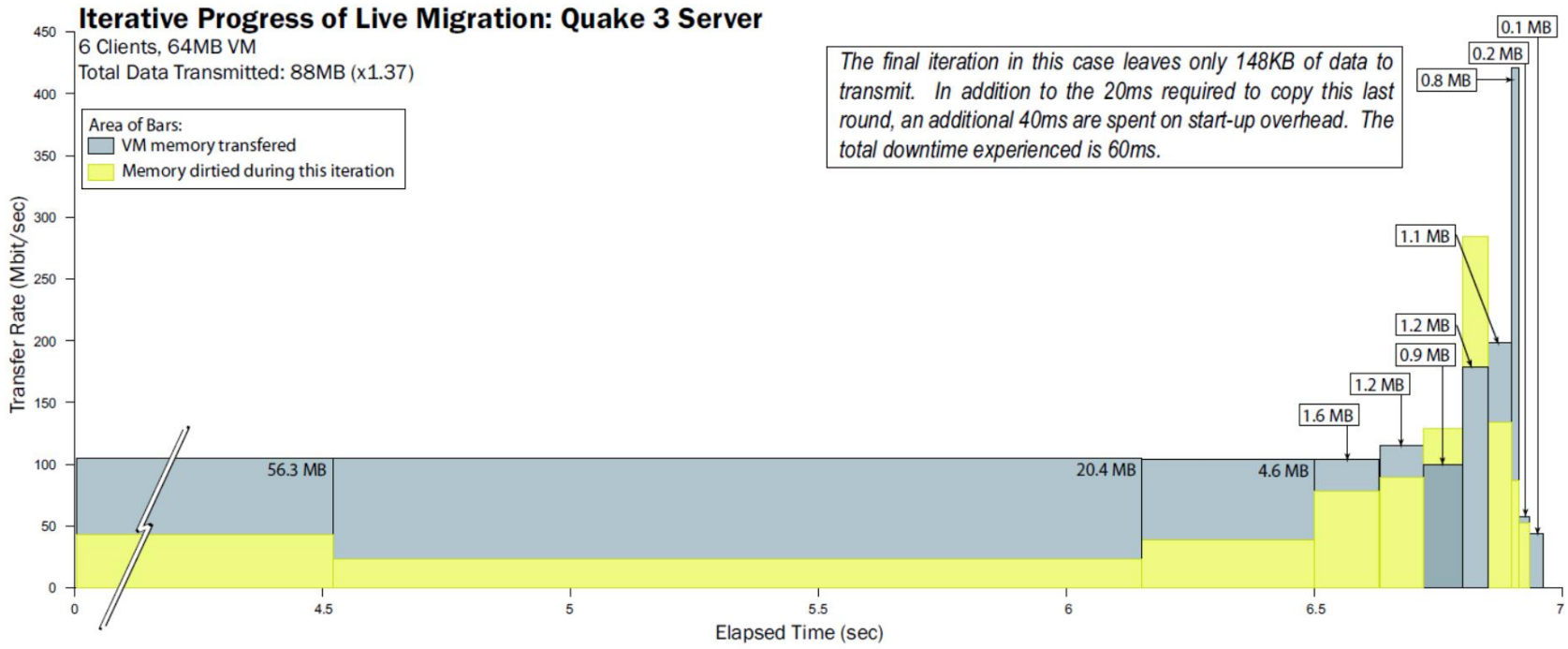
SPECweb99

- A complex application-level benchmark for evaluating web servers and the systems that host them
- Total downtime = 210ms



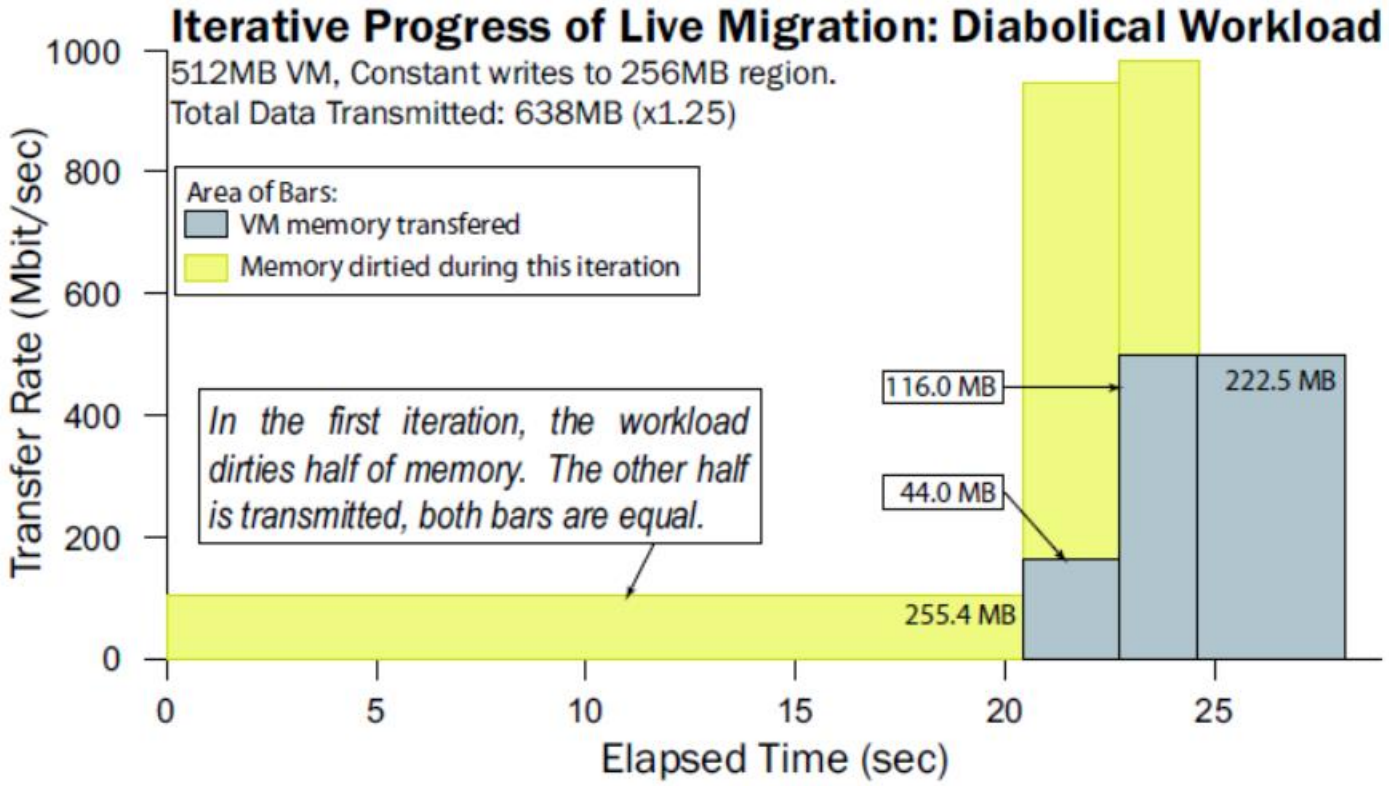
Quake 3

- A multiplayer on-line game server
- Total downtime = 60ms



MMuncher

- A VM is writing to memory faster than can be transferred
- Total downtime = 3.5 seconds



Future Work

- Cluster management
  - To develop cluster control software which can make informed decisions as to the placement and movement of VMs
- Wide Area Network Redirection
  - When migrating outside the local subnet
  - OS will have to obtain a new IP address, or some kind of indirection layer must exist
- Migrating Block Devices
  - Local disks are considerably larger than volatile memory

Conclusion

- A pre-copy live migration method on Xen VMM
- Introduce the concept of Writable Working Set (WWS)
- Dynamic network-bandwidth adaptation
  - Balances short downtime with low average network contention and CPU usage
  - Minimal impact on running services
- Small downtime with realistic server

Reference

- Christopher Clark, Keir Fraser, Steven Hand, Jacob Corm Hansen, Eric Jul, Christian Limpach, Ian Pratt, and Andrew Warfield, “Live Migration of Virtual Machines,” NSDI, 2005.