



Semi-active Workload Replication and Live Migration with Paravirtual Machines

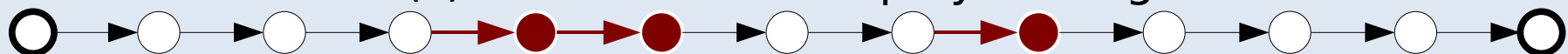
Xen Summit, Spring 07, NY

Daniel Stodden <stodden@cs.tum.edu>
Technische Universität München, Germany

Semi-active Replication



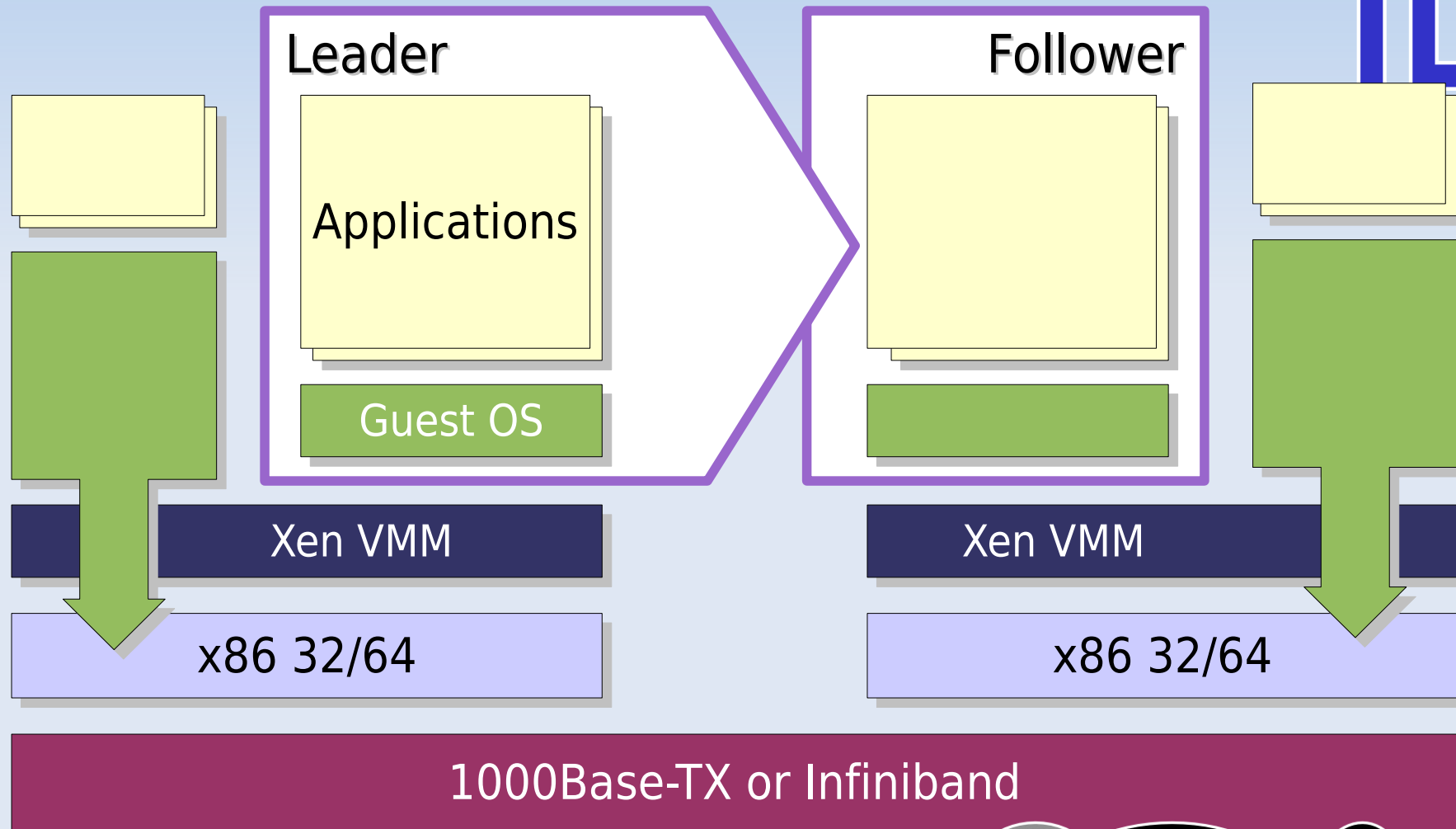
- Active Replication
 - Multiple [distributed] instances of one and the same component executing simultaneously
 - State machine approach:
 - same machine state
 - same input -> same transition, same result state.
 - But symmetric replicas require consensus
- *Semi-active Replication*
 - Asymmetric: One *leader*, one or more *followers*
 - Control *non-determinism*
 - Leader: Logs any machine 'event' affecting flow of control and data
 - Follower(s): Consume and replay the log.



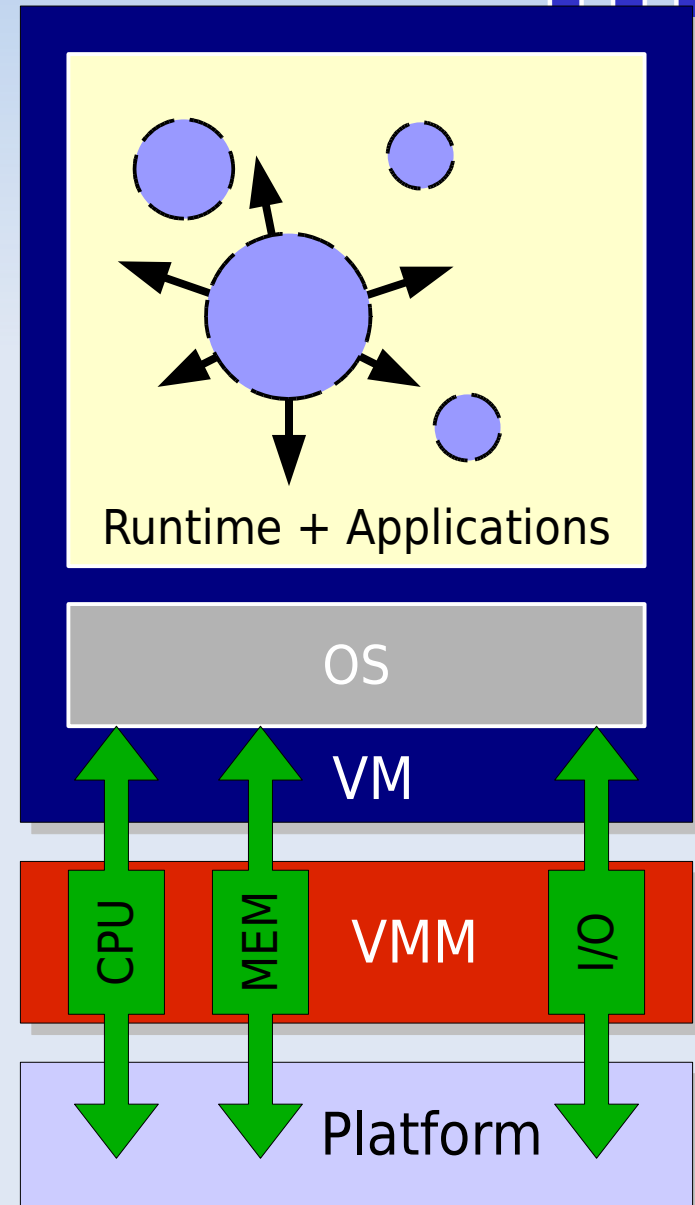
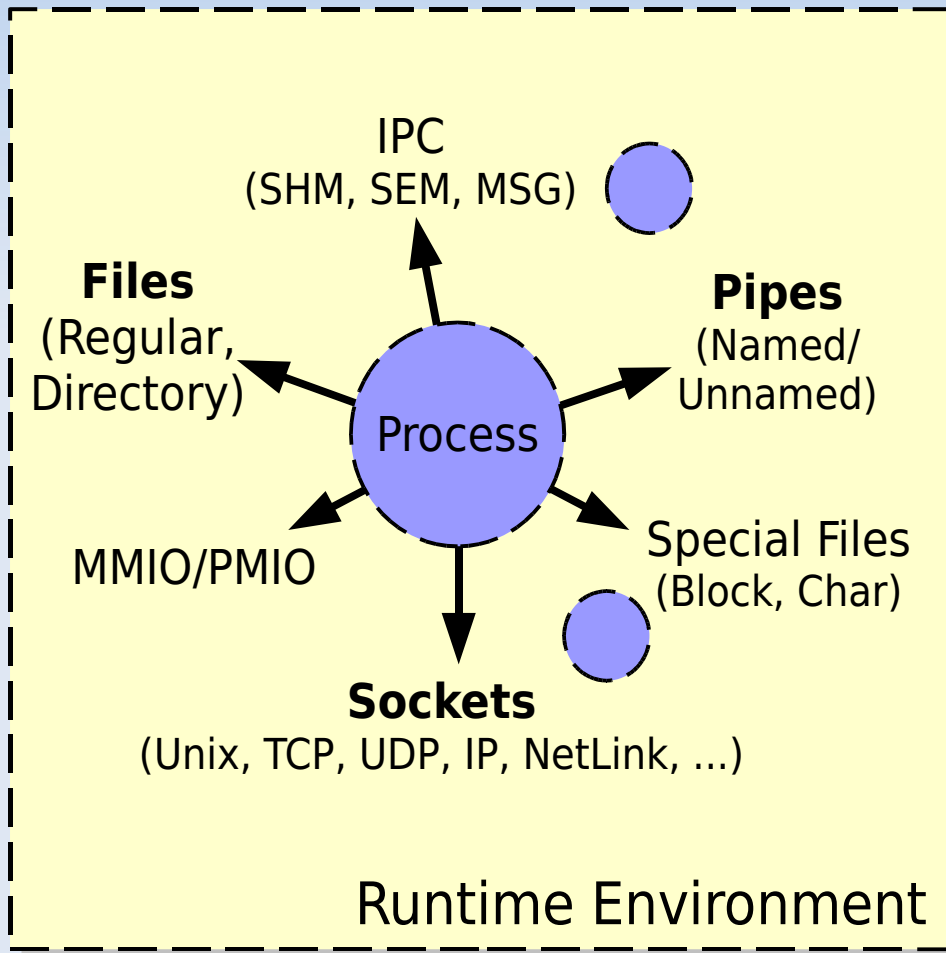
Application



- Fault Tolerance
 - Run system until it fails, and beyond.
- Debugging
 - Run system until it fails, then finally understand why.
- Intrusion analysis
 - Run system until it fails, then see .. whatever.
- ...



Why VMs?



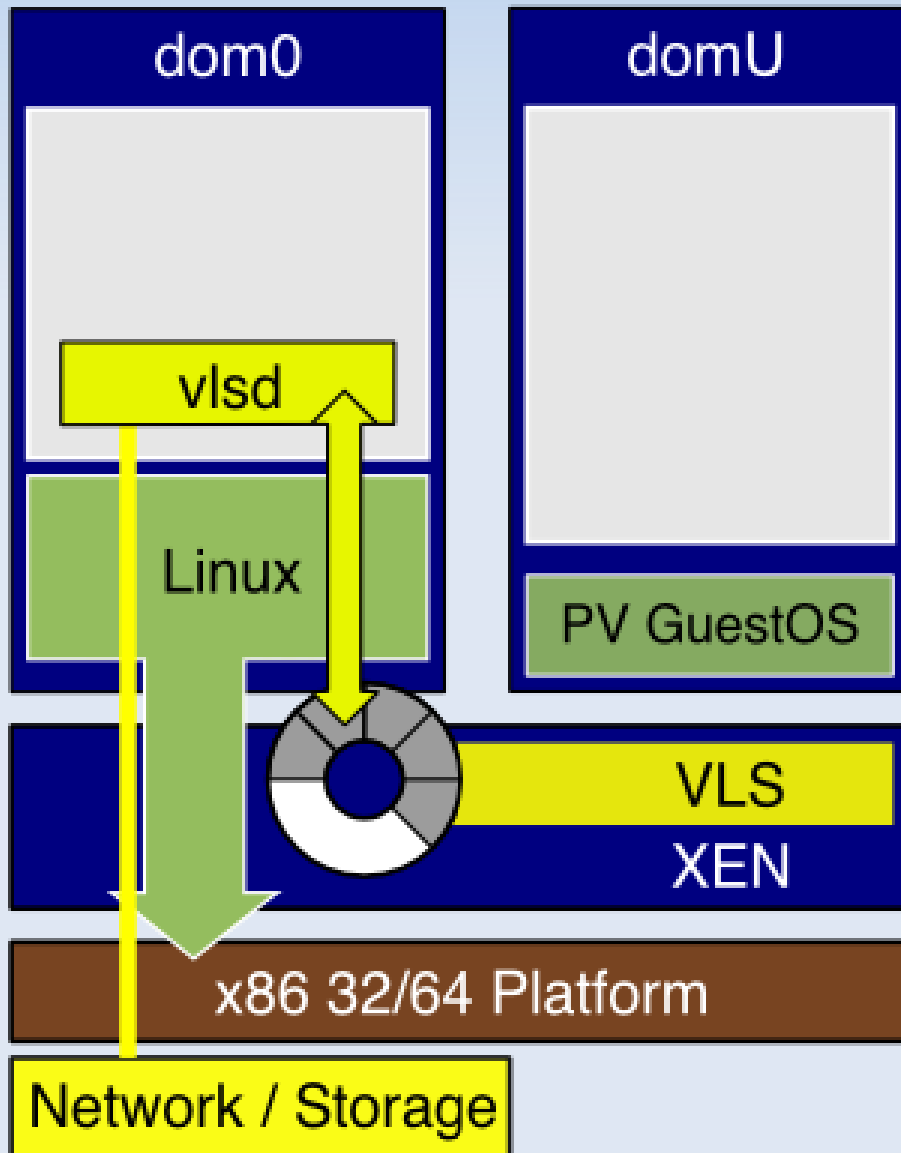
Conventional process environments are more complex, thereby much harder to control (and maintain).

VM Nondeterminism



- Synchronous
 - Driven by VM execution
 - I/O
 - Net/Blockfront drivers
 - Console I/O
 - System Time: **RDTSC**
- Asynchronous
 - Driven by interrupts
 - Require *precise* replay on followers
 - Differences in the point of delivery produce different state transitions
 - Managed with instruction counters (IC)
 - VLS: x86 Performance Counters

Xen/VLS: Architecture



- Dom0
 - Userspace control
 - XEN_DOMCTL_vls_op
 - domU ATTACH
 - domU DETACH
- Xen
 - Trace/Replay Mode
- Determinant Log Transfer
 - Single Shared Ring
 - Xen-EventChannel

Application: VM Migration



- Xen 3.0: Clark et al., Cambridge, *Live Migration of Virtual Machines*, 2005

- Two Phases

1. Iterative Pre-Copy

- Mark & Transmit dirty pages
- Iterate
- Converges: *Writable working set (WWS)*, due to ref. locality

2. Stop-and-Copy

- Suspend VM execution
- Migrate WWS
- Resume on destination host

Why does Migration matter?



- For VLS
 - Bootstrapping remote replicas
 - `xm clone ...`
- *With* VLS: Maybe aid Migration.
 - Potentially solve some corner cases
- Xen Stop-and-copy:
 - Typical service outage: 50 ms (Quake) – 201ms (SPECweb)
 - Breaks with excess memory bandwidth consumption

Demand Migration



- One alternative: *Demand Migration*
 - Transfer execution to the destination node early
 - Fetch missing pages upon demand
- Observations
 - Page transmission order respects processing demand
 - Stop-and-copy did not
 - But transfers are synchronous!
 - Request initiates on destination node
 - Takes a round trip to answer
 - Many missing pages to *block* on

Semi-active Migration



- Proposed algorithm
 1. Iterative pre-copy
 - Converge to WWS
 2. Switch to "Semi-active mode"
 - Activates destination node as a *follower*
 - Source continues as the *leader*
 - Running ahead of destination node
 - *Demand* transfer of missing pages
 - Terminates upon memory transfer completion
- Expected properties
 - Similar to demand migration
 - But ahead-of-time page transfer
 - Full Service continuity

Project Status



- Mini-OS
 - Traced & Replayable
 - Wallclock and System Time replay
 - RDTSC Simulation
 - VCPU/Poll/Interval Timers
 - Verbose Mode (Debug Support)
 - Hypercalls, Events pending, ...
- Linux domU currently work in progress
 - Block / Network / Console I/O
 - Coming RSN. Well, hopefully
- Repo? <http://somacoma.de/~dns/share/hg/xen/v1s>



Thanks for listening.