Can Deterministic Replay be an Enabling Tool for Mobile Computing?

ODR: Output-Deterministic Replay for Multicore Debugging

Lifephone: Deterministic 1 Hosting in Smaller P

Scalable Deterministic Rep a Parallel Full-system E

Deterministic Replay: A Survey

Deterministic Replay: A Parallel Full-system E

Efficient Deterministic Replay Using Compiler

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:

Decoupled Store Completion/Silent Deterministic Replay:
Number of systems supporting replay: 0
Determinism

CPU is deterministic
Determinism

CPU is deterministic

Interrupt, device I/O

Execution history
Recording

- Ability to interpose on external inputs
- Determinism of the execution environment
- Instruction-accurate positioning of events
Why is it hard?

- Complex interposition boundary
- Concurrent reentrant environment
Why is it hard?

- Complex interposition boundary
- Concurrent reentrant environment
- Complex instruction counting logic
- Subtle divergence bugs
Why would you really care?

- Execution analysis
- Automated debugging
Abstractions for Practical VM Replay
A three-part model
A three-part model
A three-part model

Deterministic Environment

Replayed System

Non-deterministic External World
Event types

- Deterministic updates

- Non-deterministic External World

- Replayed System

Hypercalls, exceptions
Event types

- Deterministic updates
- Synchronous

Deterministic Environment

Replayed System

rdtsc(), in()

Non-deterministic
External World
Event types

- Deterministic updates
- Synchronous
- Asynchronous
Event types

- Deterministic updates
- Synchronous
- Asynchronous
- Dependent
How does it map on an actual system?
Interposition functions

```c
int trace_<function>(...) {
    event_t event = {<EVENT_TYPE>, ...};
    if(replayed_guest()) {
        if(synchronous(&event)) {
            // request replay of a specific event
            replay_current_events(..., &event, &ret);
            return ret;
        }
        // asynchronous event: suppress the
        // update but replay "optional" events
        replay_current_events(...);
        return OK;
    }
    // Pause all virtual CPUs
    pause_vm();
    trace_event(, ...);
    // Emulate original event
    ret = <original_function>(...);
    unpause_vm();
    return ret;
}
```
Lightweight interposition and logging

- Logging code is on the guest’s critical path
- A typical exit to Xen is 4000 cycles
  - Replay budget of 10-20% (or 400-800 cycles)
- Offload everything from the critical path
Lock-free, no-allocation, zero-copy
Device determinizm

- Determinizing proxies
Split devices and shared rings
Devd: general device interposition
Instruction-accurate event positioning

```
label: ... 

mov
shr
movsl
test
jne label

... Position = {EIP, branch #, ECX}
```

- Number of instructions since boot
  - Intel has a hardware counter
  - It’s not accurate
Execution scheduling

Event scheduling types
• Asynchronous
Execution scheduling

Event scheduling types

• Asynchronous
• In-place
Execution scheduling

Event scheduling types
- Asynchronous
- In-place
Execution scheduling

Event scheduling types
• Asynchronous
• In-place
• Optional
• Non-replayable

Retyping async events as synchronous
Replay on exit to guest
Every step is accompanied with a debugging mechanism

- Automatic analysis of nondeterminism
- Page guarding

Guest

Xen

Time

Create a list of protected pages
Unprotect pages on exit to guest
Reprotect pages on exit to Xen

Guest's writes are OK

write()
Every step is accompanied with a debugging mechanism

- Automatic analysis of nondeterminism
  - Page guarding
- Off-line comparison of execution
  - Intel branch store trace (BTS) facility
    - Record all taken branches in a memory buffer
  - Support for resolving symbols
Every step is accompanied with a debugging mechanism

- Automatic analysis of nondeterminism
  - Page guarding
- Off-line comparison of execution
  - Intel branch store trace (BTS) facility
    - Record all taken branches in a memory buffer
    - Support for resolving symbols
- Run-time comparison tools
  - Compare guest’s state between original and replay runs
- Trace from all parts of your system
  - Xen, Domain 0, Guests
- Support performance tracing
  - Xentrace messages
How much overhead?
32-bit x86 PV-guests
xen-unstable near v3.0.4
  We rely on working shadow page tables
1-CPU time-traveling guests
  No SMP replay
  Dom0 and Xen are SMP of course

Test machine
  4 cores
  1Gbps network
  130 MB/s disks
CPU-intensive workloads

% of normal Xen

- gnugo.tactics
- ogg.tibetan-chant
- bzip2.decompress-9
- scimark2.small
- ddraw.d300
- povray.reduced
- openssl.aes
- bzip2.compress-9

XenTT
Systems workloads (Phoronix)
Network throughput

![Bar Chart]

**MB/s (higher is better)**

- **TCP Send**
  - Xen: 110 MB/s
  - XenTT: 100 MB/s

- **TCP Receive**
  - Xen: 110 MB/s
  - XenTT: 100 MB/s

Legend:
- **Xen**
- **XenTT**
Network delay

- Ping
  - Idle
  - 100MB/s stream

- Xen
- XenTT

- ms (lower is better)
## Log size

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Compressed (gzip)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linux boot</strong></td>
<td>903 MB</td>
<td>145 MB</td>
</tr>
<tr>
<td><strong>Idle overnight (12 hours)</strong></td>
<td>2 GB</td>
<td>529 MB</td>
</tr>
<tr>
<td>Growth rate</td>
<td>167 MB/h (4 GB/day)</td>
<td>44 MB/h (1 GB/day)</td>
</tr>
<tr>
<td><strong>TCP receive (4 GB stream)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event log</td>
<td>1.8 GB</td>
<td>342 MB</td>
</tr>
<tr>
<td>Payload log</td>
<td>4.4 GB</td>
<td>Payload dependent</td>
</tr>
<tr>
<td><strong>Disk write (4 GB file)</strong></td>
<td>600 MB</td>
<td>145 MB</td>
</tr>
<tr>
<td><strong>Disk read (4 GB file)</strong></td>
<td>414 MB</td>
<td>62 MB</td>
</tr>
</tbody>
</table>
A tribute (related work)

- ENIAC (1948)
- Exdams (1969)
  - The first record-replay system
- ReVirt (“I’ll never work on it again!” [ReVirt team member])
- VMware (“They did it, but it was a lot of work.” [VMware team])
- QEMU
  - Lots of simplifying assumptions
- rr Mozilla
Principles (recap)

- Three-part model
  - Event types
- General interposition functions
  - Simple locking model
- Determinizing proxies
  - Active messages
- Fully-asynchronous logging
- Debugging tools for scaling development
Conclusions

- Principles are important and will help future implementations
  - Xen is arguably the hardest environment to implement replay
  - KVM + QEMU should be easier

- I/O delay goes up
  - You can address it with more cores
  - But still cache-coherence has a non-zero cost

- Branch counters are fragile
  - Our code works on several server CPUs
  - Fails on a laptop with the CPU from the same model/family line

- Mainstream replay is possible for us as a community
Thank you.

More questions: aburtsev@flux.utah.edu