XenSummit

XenTT: Deterministic Systems Analysis in Xen

Anton Burtsev, David Johnson, Chung Hwan Kim, Mike Hibler, Eric Eide, John Regehr

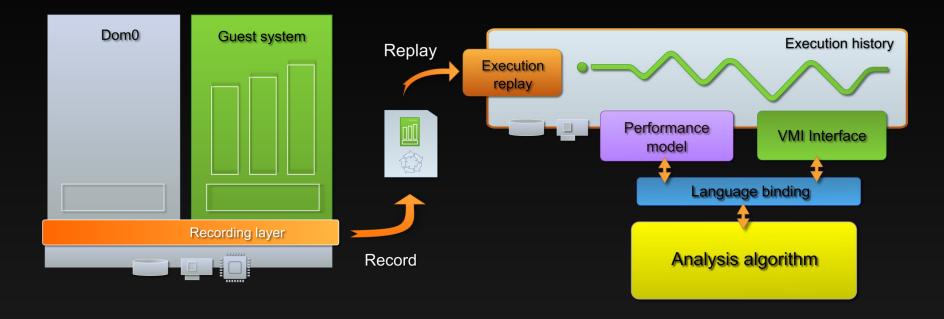
University of Utah

August 27-28, 2012 San Diego, CA, USA



- Record execution history of a guest VM
- Recreate it in an instruction-accurate way

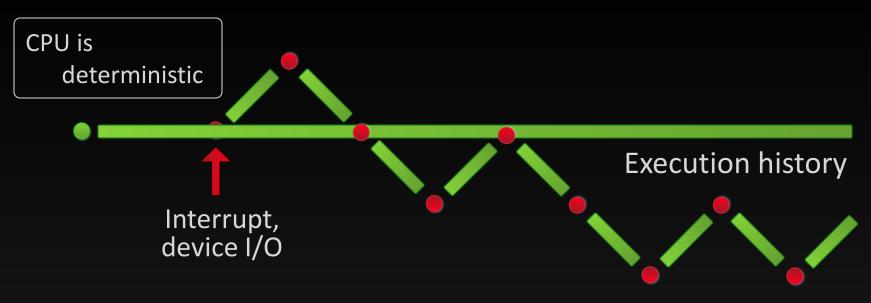
							and the second sec
		Keywords Hulti promer Bipha, 10/7 when Sharol W	The and Separated Constrainty (SC) execution. Moreover, DeCores See and exact space 1.0, of the other constrainty of the constraints of the constraints (11) of an a proper (1, 11). The fair, fairy and in the constraints of the constraints		and a second sec		get_tree_pag
The Deterministic Exe		excises cells dentes rections to: L. Introduction	In Proceedings of the Desmontened Symposium on Computer Architecture, June 2013	and the second s	CONTRACTOR OF CO	t chourt	zone
How Well Does it Actua		tion tracts Alaho tongka upropuerts monthine has a			States and a state with the states	CSGOCT	
Tim Bergan Joseph Devietti N	ick toctory research. Unfortunately, existing trace collection techniques are often slow (due to software tracing over the software tracing over		BugNet: Continuously Recording Program Execution for	A REAL PROPERTY AND ADDRESS OF AD			alloc_pages
Childrenity of Wads (thingan divident), shared, lawroop)			Deterministic Replay Debugging				page / /
Abstract	quiremente). Regardless of the method of collection, de 5 tailed trace like are generally large and inconvenient to	in surprising) uni prutus socurs. Howeve, with an off- the most of cost-cold recert a malaric program occording		10000		Ch'an	page /
This paper takes a critical look at the benefits provided by	inter and that	an drive that exception using any haves wight d the their an strength of the producting in [5]. If the recording is the set of the replacing in [5]. If the recording is	Satish Newymannesty Giller Polines Bred Calder	A REAL PROPERTY AND ADDRESS OF ADDRESS ADDRESS OF ADDRESS OF ADDR	A COLORED TO A COL	rtirg	
interest file at deterministic succession techniques. Specifically, we look at finar applications of deterministic encou-	We present R/Durse, a trace collection load based on the	an checkengers meaned. In the sense of a software factors, r	Department of Computer Asiancer and Engineering University of California, San Dany		Concession of the Area and the		and many training transferd
tion: defugging, fault-tolerant replacation, testing, and se-	(k) deterministic riphey technology of the VMware hyper- visor. Boltram surgers in two starts: contarious and provident.	information these study durings and today.	(sorted), generation, califier / sizes record who	2 Objection of the local diversity of the loc	A REAL PROPERTY OF A REAL PROPER		get_page_from_freelist
carity. For each application, we docum what an ideal sys- tem would provide, and then isok at how desemblishing can-	den reponsion. Er/Daor capturing accumulator the raini-	 Bowerse while date of the option of all is and second or chosen or is shall obtain the model. 			A REAL PROPERTY AND A REAL		try to free pages
terms commune to the ideal. Purther, we discuss adarmative an-	The statement of information necessary to later respondences in more detailed economics mass. It expenses (records)	plote 1 entrand or if adhead or and or a sense are adhead plote 1 entrand or a set in the sense are adhead are set in the appendix entrand or and a set in the sense of a set indicated or any entrandom or or any	Bennet present/for factors have	A REAL PROPERTY AND ADDRESS OF			
or not these alternatives are more suitable. Along the way,	andy non-deterministic events resulting in low time and	al solution MERCINE Database (2-in 2001 New York, 502 100		A DE LA DE LA DELLA DELL	A CONTRACTOR OF A CONTRACTOR O		arch/i386/mm/ /
we identify open questions and suggest faster work. Ultimately, we find that there are competitive alternatives 2.	pass overheads (as low as 5% run-time overhead, as low as 0.5 here per thousand instructions by growth rate)		addresses and a dark in compared sole. To sum develop- ter, se popular to darket sole to compare a sole to	A REAL PROPERTY OF THE REAL PR	The PLANT AND A STREET	nifie .	nhysical momony
			Expension on productors rate. The adjoinance colored to: free the cruck of a program can be used by the developer work.				physical memory
programs, that determining has high, drough approven, po- terminative reprive testing, and that deterministic has deduced in			ing in fair marities processes to introduct all rights for a the shift's to open entrop to entry the open and the state that		And the second sec	Internet	onerations
		re gran bring traord. For example, traor collection		A CONTRACTOR	The second	incerrupt	physical memory operations
Partnesses, determines is a setted solution for all four applications, this confers a distinct advantage over point in-	near of VMware Workstation 6.2 currently sealable in Wis- down and Lines Ranses for corresponding 13/27 adultance	in a memory trace. Perhaps more suffily, tracing	support of the second s		The second se		die
intom that do not compose well with one another. The	rea No special tracing hordware is required.	should minimize time dilution and memory dilution which occurs when tracing causis a program to run	to program to exercises, they be interest on heavy date of the registry of the second to provide relation of the librate to star, but not the provide relationship to deard second to provide relationship to		Con Dia Management	-	/ y''e
I. Introduction all	gen We have three key results. First, we find that trace ed-	abover (him its own perspective) or to commi-	sparsing cuses. But not approach provide the ability in re-	A DESCRIPTION OF A DESC		switch to	
There is a conservery in the community that wondeterministen idea	B.V Second American's such as dealers in the second					and and and	
rades the desolopment of parallel and concurrent software and substantially ment difficult Recent work has proceeded in-	Kall compensation large triage film. Third, performing the trace	· Robusing trace fits size. The trace sufficient tax	FO, superspit and DAX backets water and share obspitality water for the second and the second state and an experiment of the second state of th				
indag reedstantions drough dearmineast consults (7, car		re should produce highly compact trace files, which are major to store and share with other researchers	does accepter a final over deep of the oppose space (or oppose). You, while a sectory of tender and system can relative, the observation of the sectory of the sectory of the last rest of the sectory o				
5-7, 15, 16, 20; Proprients of deterministic excendes as the our that it stratifies many challenging today faced by well, for	epitem (user/supervisor level, CPU, periphenal devices)	Compact trace files also enable long execution trace	and back as the developer to be a set of the				
ware developen. This paper examines how mach determin. The initia parameters actually samplifies from challenging tasks. will	ne : BrTrues is a rapidly coolsing technology. We would like	to be collected for different analysis needs.	1 Introduction on deterministic registring the meta-chain encoded in one				
We expense four specific tasks: debuscing (Section 3), any	378 of Er/Tree is constant or militation manuals to help or	 Fast, inexpensive and may to operate, 			And a state of the		
fault sciences replication (Section 4), assing (Section 5), and an transval of covert characteristic for security (Rection 6). For each data		To the front of our knowledge, more of the manting trust		A CONTRACTOR OF A CONTRACTOR O			
task, we ask 16 w much-does deterministic stopping the task? yest	37.6	collection techniques sufficiently solves all these clud longes. We latisfy server these techniques in Navian 6		A REAL PROPERTY AND ADDRESS OF	COMPANY OF THE OWNER.		
What are discritative approaches, not involving discriminism, rep- and how dires determinism compare to those administrees?		tanges. We failedly survey these heatroques to Nectors 5. This paper presents a deterministic rephy hand cares	2001 (4) Understands, subsure regioners part verdination in general determine and review colls, and fair its same part- teringenes are required, and a software review rever after as designed approved Quality Assessment in Norma (a James Party), and the Same are an application test tags.		A DESCRIPTION OF A DESC		
An the benefics of deterministic worth the cost? Our posts are to identify weak spots in the current state of the art, then	sk)	tion trace technique, Rel'suer, which hotter solute al- ing these challenges. Rel'suer decomposes traditional trace	despeng typeses Quddy Assesses is bound to have been in				
		fig these channenges. For Frace decomposes therefore an in collection process into two steps: fo Proce constanting and	te deminde software perdan, and Eas her mersenanged omder: te dem mere upd minnes relabilite, which sendre in relevant. Microsoft Dr. Timmarted [2] and Monika's Talibards [7] and	A REAL PROPERTY AND A REAL	Constant and Constant		
most important open problems. Kay among these problems date is understanding the performance tents of descretations. date							
In the root of this paper, we first provide context and dry	1786		Tocking down and thing logs in released or frome on her. Here program stock. All does tools codest administration that rep- e sublicant, contain a significant margin of tane and margin - owner, the fluid superbolt of execution refer when the accurate				
summaries the current state of deterministic execution close tion 25. We then examine the henedits and costs of draw tion	pen D. O. an and a second		The same difficulty law is being olde to opposition the log of the simulation While these conductyments have some while; it is highly developer time. More of the startest defaustors review on a single deduction of the startest defaustors		and the second s		
ministic success, as applied to four applications (Sec-	icetze-sif delartezza.		core doup [R. 7], viack represent the faul time of the cystem. antinetisso-executed before the creds. An example of soch a se- plener the creds. Undertaurable, this initiation has been workfully a form in them. Research and Die Konschen 1107. There built a soft-	THE TAXABLE PARTY OF TAXABLE PARTY	Contraction of the second of t		
tows 3-61, documents the performance limits of determinism (Section 7), and finally conclude (Section 8) ing	Lockslep quanta an a common approach for provid- strong determining. The klea is to divide execution into				A DECEMBER OF THE OWNER OWNER OF THE OWNER	8	
			non-schoolly sequentials for the program to end up with the state on go back in tase and repeatedly in-micros the program				
			1		A REAL PROPERTY AND A REAL	1.4	
					A DECEMBER OF THE OWNER OWNER OF THE OWNER		
				And and a state of the state of	A DESCRIPTION OF TAXABLE PARTY OF TAXABLE PARTY.		
						21	
					Contraction of the second seco		
					AND ALL AND A REAL PROPERTY AND A REAL PROPERT		
				Call and a little second	A DECEMBER OF A		
					No. of Concession, Name		



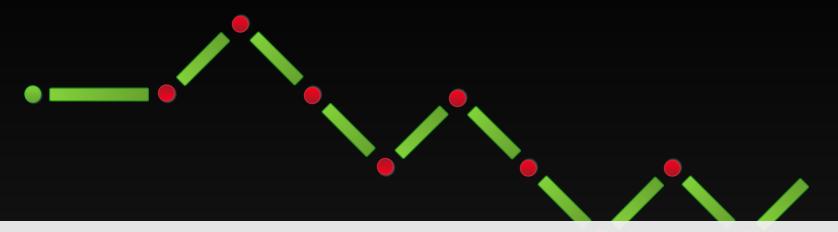
- Execution replay is a right way to analyze systems
- We need a practical tool!

Deterministic replay

Determinism



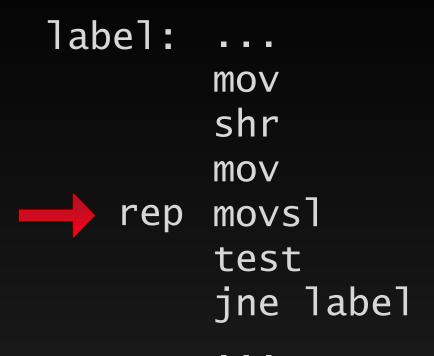
Recording



- Determinism of the execution environment
- Instruction-accurate position of events



Instruction-accurate position of events



- Number of instructions since boot
 - Intel has a hardware counter
 - It's not accurate

- Hardware instruction counter
 - Preempt execution of a system at the same instruction
 - Hardware instruction counter + single-stepping

Determinism in Xen

Nondeterministic events

Simple Model

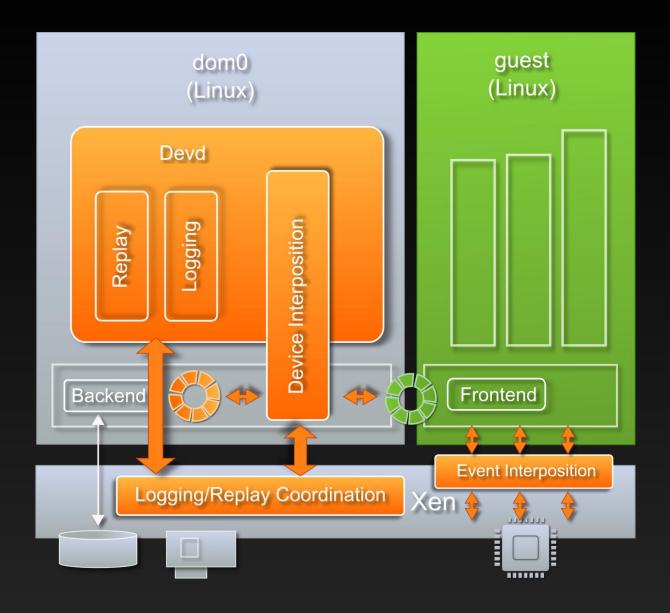
System = memory pages + registers

Events = memory updates (time, device I/O, system calls)

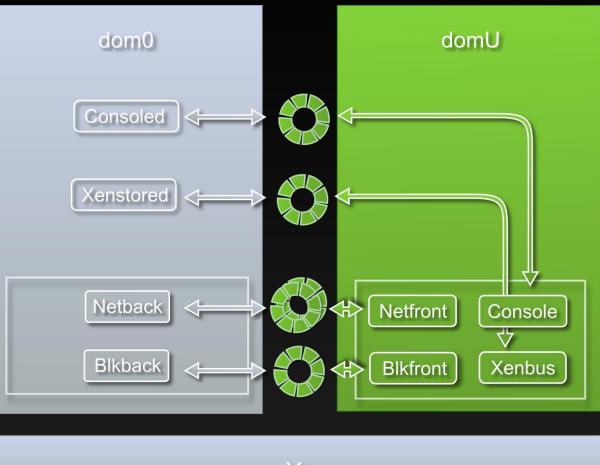
Control flow updates = registers + stack (interrupts, events)

Some examples

- Instruction emulation (e.g. cpuid, rdtsc, in/out)
 - Return the values of the original run
- Hypercalls
 - Re-execute to ensure determinism of the hypervisor
- Time
 - Shared info page + rdtsc
- Exceptions
 - Deterministic, re-execute
- Interrupts
 - Force re-execution of the interrupt frame (bounce frame) code in entry.S
- Shared info updates
 - Replay original values
- Memory
 - Shadow page tables

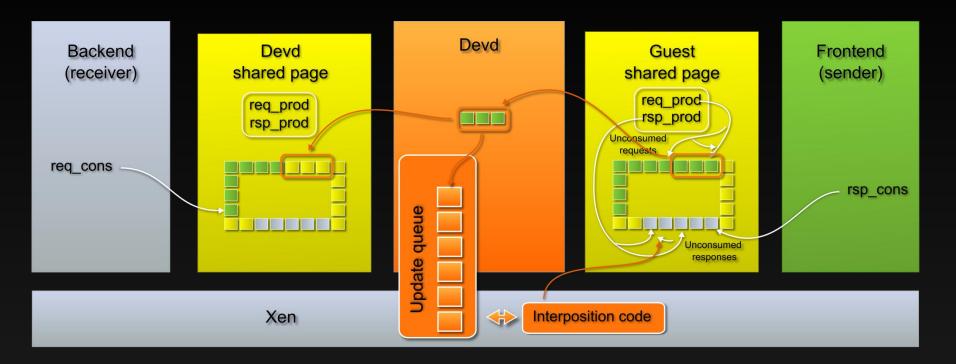


Xen devices





Device interposition

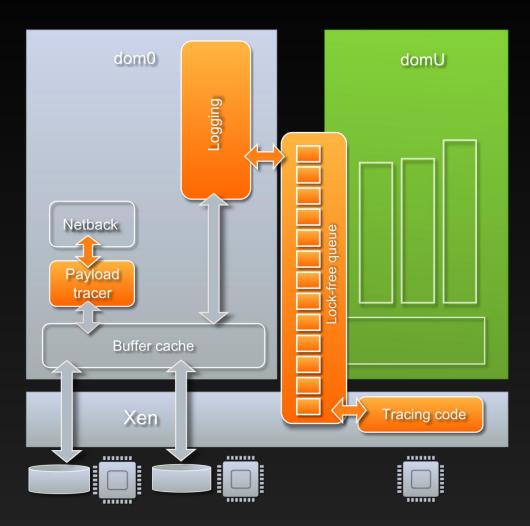


Devd ensures determinism of updates to the guest's shared ring buffers

Replay touches many parts of Xen

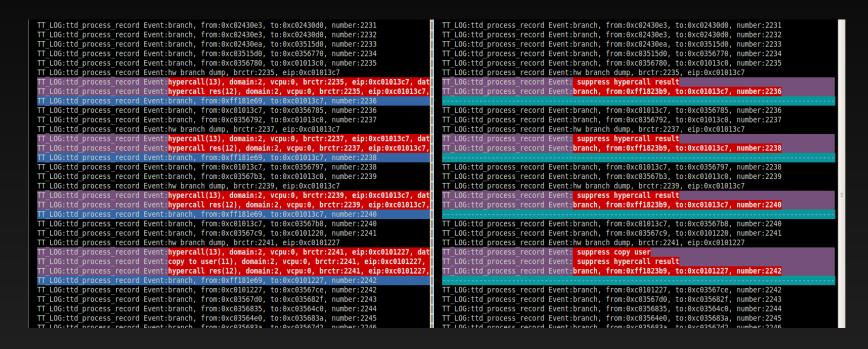
- Device discovery
 - Devd implements a concept of a device bus
 - Discovers new devices in Xenstore
 - Binds new devices with drivers
- Xenstore transactions
 - During replay, transactions from replayed guest can't fail
 - They will not be re-executed
- Out-of-order device responses
 - Disk and network responses can arrive out-of-order
- Disk logging
 - Disk payload is deterministic
 - LVM snapshots
- Network logging
 - In-kernel logging of the network payload

Low-overhead logging



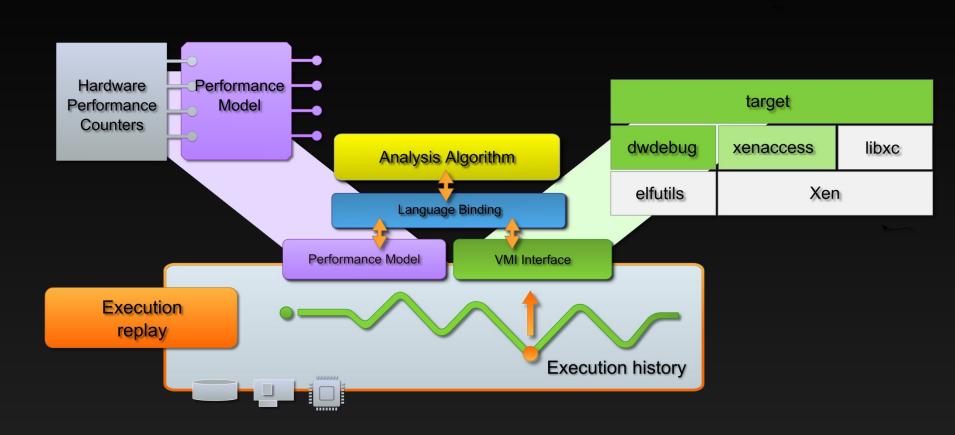
Are we sure that executions identical?

- Intel branch store trace facility
- Record all taken branches in a memory buffer
 - Compare original and replay runs

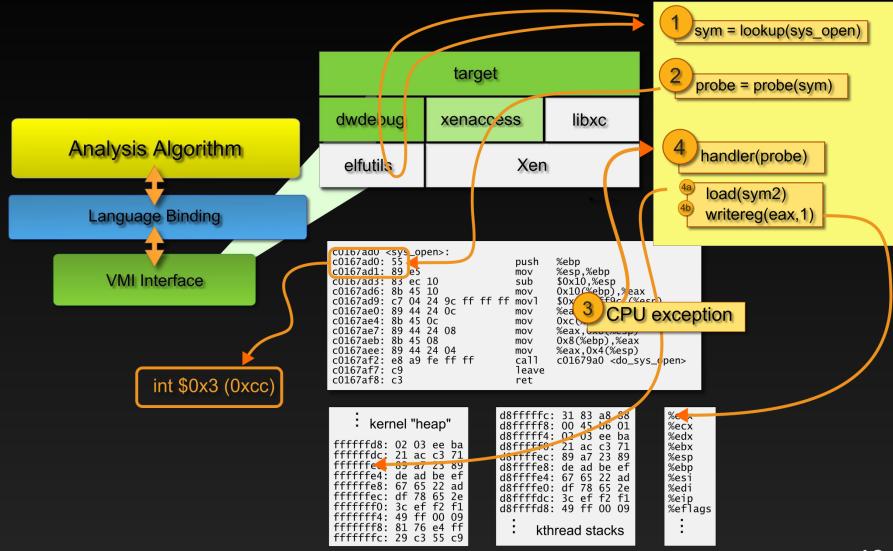


Analysis Engine and Virtual Machine Introspection

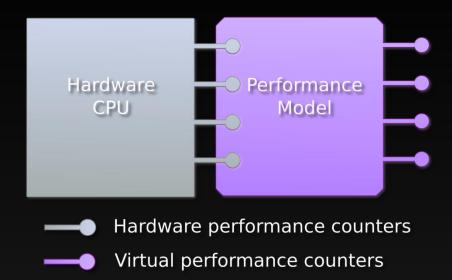
Analysis framework



Virtual Machine Introspection (VMI)



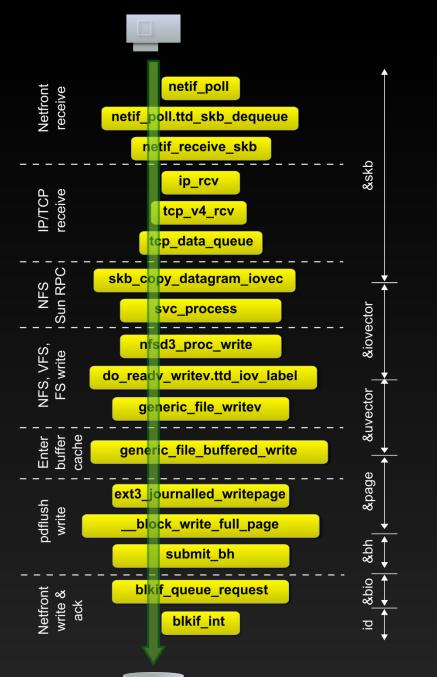
Performance model



- Account for effects of replay
- Translate performance between original and replay runs
- Re-execution approach to performance

Virt cntr = Virt cntr_{start} + Δ (Real cntr)

Analysis Examples



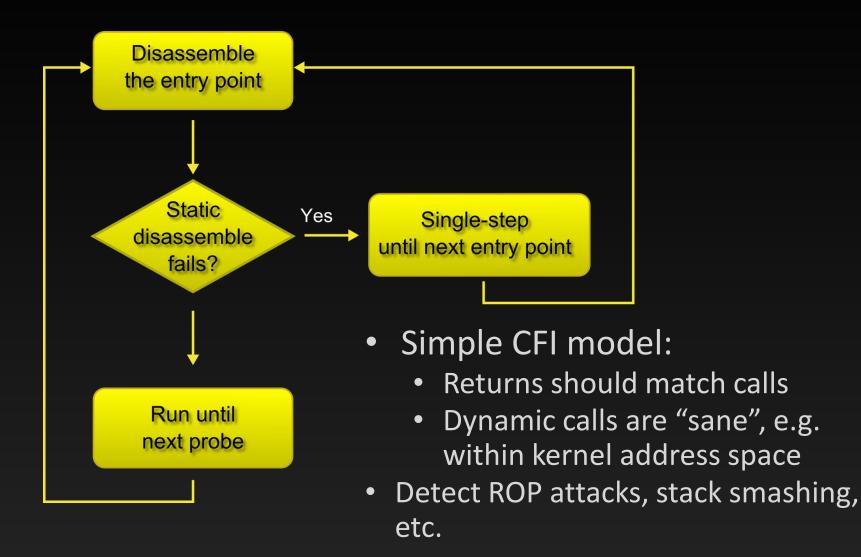
NFS request processing path

- How much time requests spend in each subsystem?
- Request tracking
 - Address of the kernel data structure is a unique identifier
 - Join identifiers when requests move between subsystems

Execution context tracking

- Execution context
 - Context switches
 - schedule.switch_tasks
 - User/kernel
 - System call transitions
 - system_call
 - Interrupts and exceptions
 - do_IRQ
 - do_pagefault
 - do_* (divide_error, debug, nmi, int3...)
- Make analysis context aware
 - Filter probes by context, e.g.
 - All pagefaults from the process "foo"

Control Flow Integrity (CFI)



Execution trace

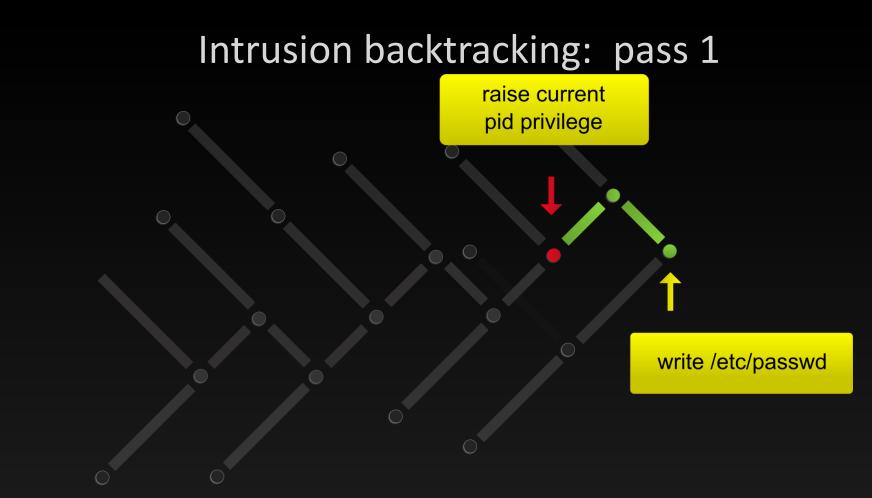
sys_sendfile do_sendfile fget_light rw_verify_area fget_light rw_verify_area shmem_file_sendfile do_shmem_file_read shmem_getpage
find_lock_page radix_tree_lookup shmem_recalc_inode shmem_swp_alloc shmem_swp_entry kmap_atomic kmap_atomic page_address kunmap_atomic find_get_page radix_tree_lookup file_send_actor sock_sendpage UNKNOWN FUNCTION (addr:0x0000000)

- CFI records a trace of function calls
 - sys_sendfile is the last system call before control flow jumps to 0x0

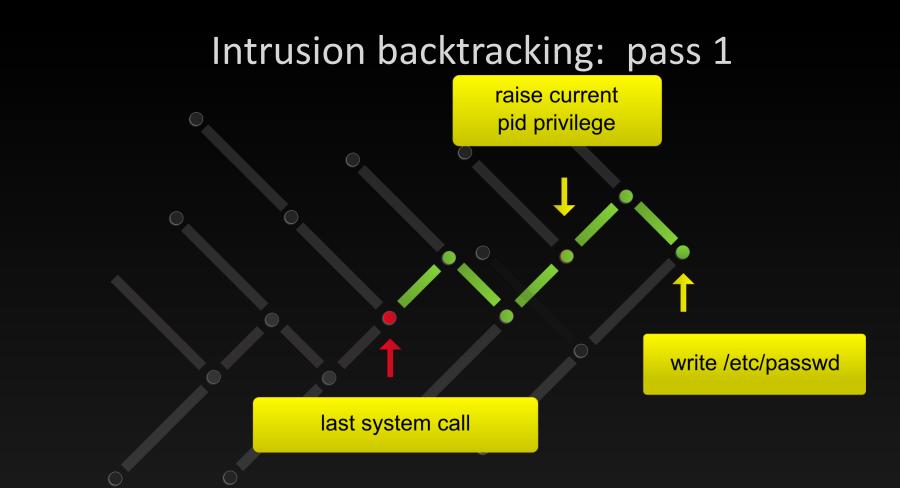
Intrusion backtracking

write /etc/passwd

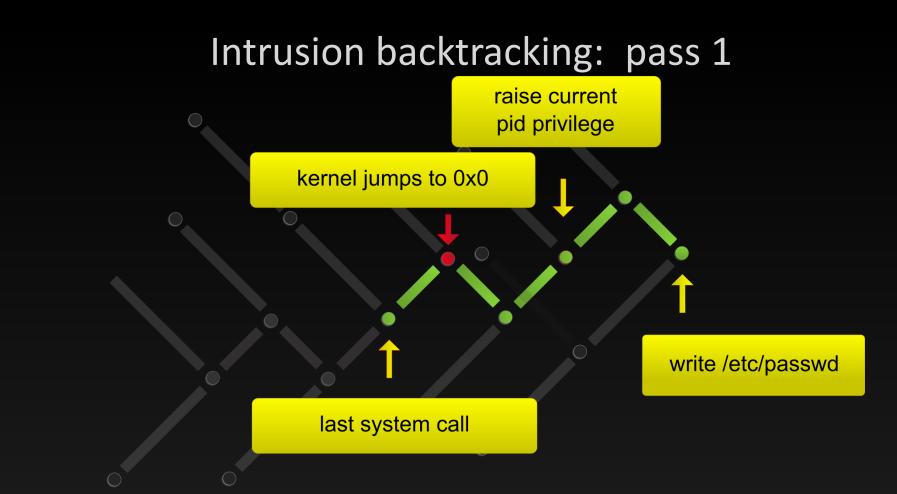
- Track accesses to "/etc/passwd"
- Probe sys_open
 - Filter by file name
 - Find process ID, branch counter



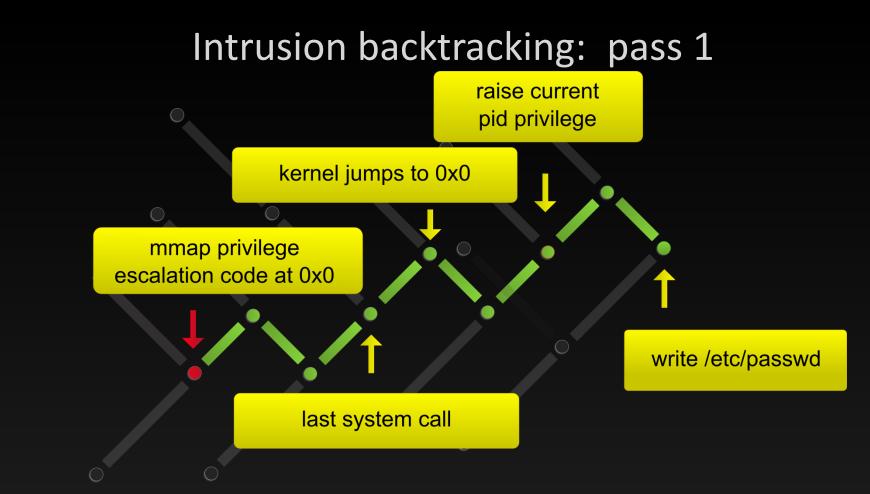
- Process or it's parent escalated privileges
- Watch write accesses to &task->uid
 - Filter by parents of the offending process



- Find the syscall inside which privileges are escalated
- Probe sys_* all system call entry and exit points
 - Filter by the offending process ID



- Privilege escalation is a CFI violation
- Start CFI analysis from the last system call
- Find %EIP at which CFI is violated, and location of the shell code (0x0)



- Find at which point address 0x0000000 gets mapped
- Probe do_page_fault

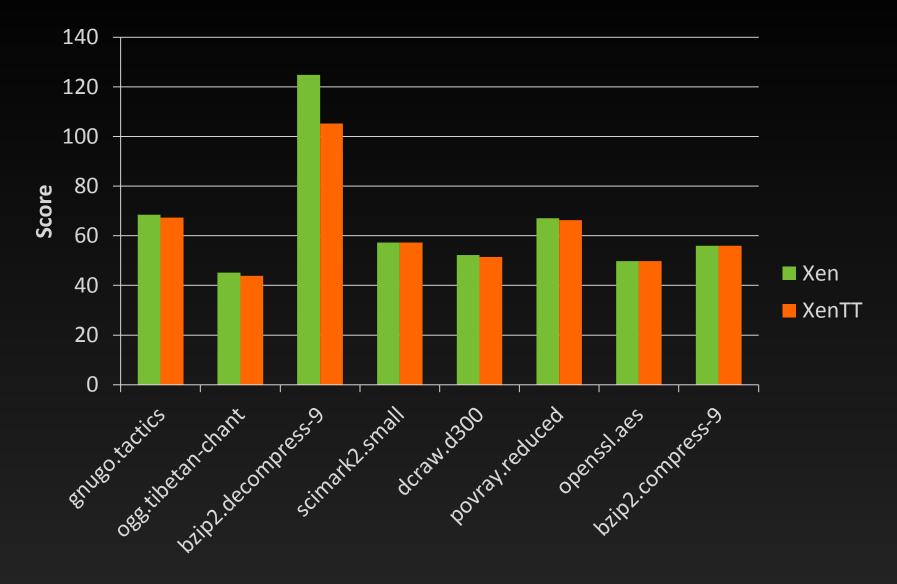
More mechanisms

- Execution traces
 - BTS trace of all taken branches
 - Instruction traces
- Memory (variable) access traces
 - Intersect with the execution trace
 - See where variables get accessed

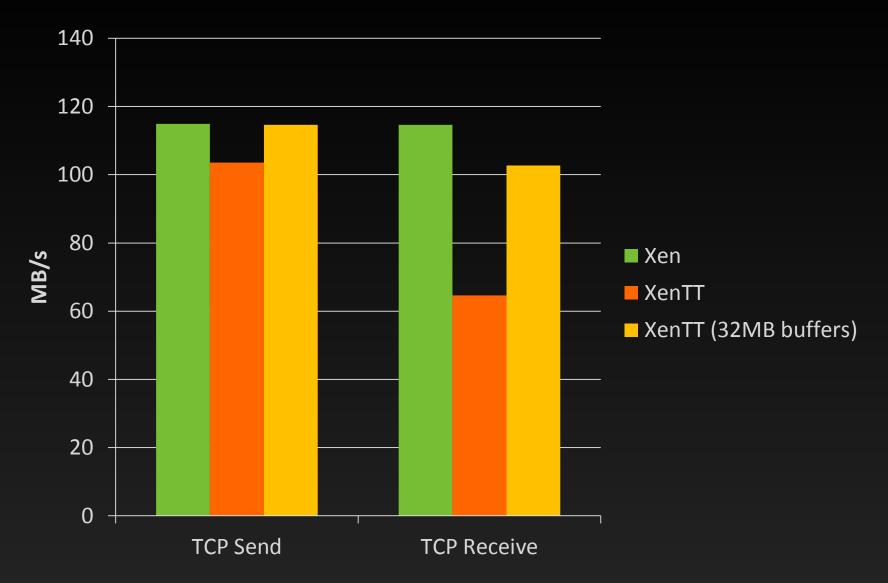
How much overhead?

- 32-bit x86 PV-guests
- xen-unstable near v3.0.4
 - We rely on a 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

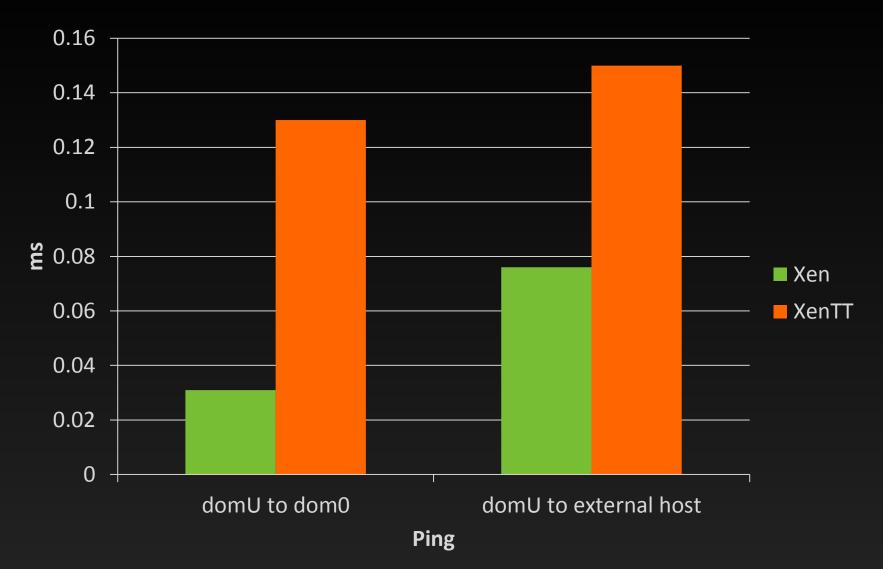
CPU



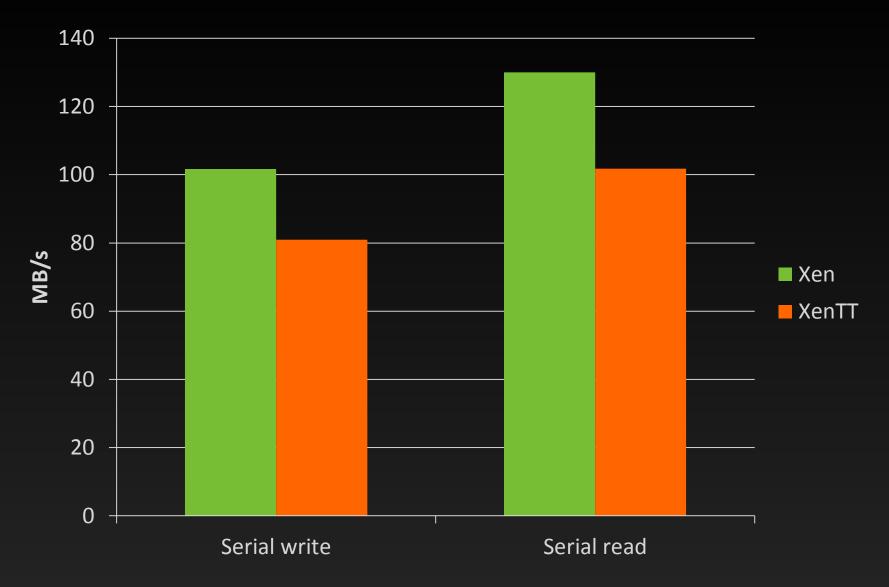
Network throughput



Network delay



Disk I/O



	Raw	Compressed (gzip)
Linux boot		
Event log	4.3 GB	0.8 GB
Idle overnight (14 hours)		
Event log	6.2 GB	1.6 GB
Growth rate	440 MB/hour	114 MB/hour (2.7GB/day)
TCP receive (1.63 GB data stream)		
Event log	1.76 GB	342 MB
Payload log	1.79 GB	Payload dependent
Disk write (4 GB file)		
Event log	1.8 GB	350 MB
Disk read (4 GB file)		
Event log	0.59 GB	

Lessons learnt

Scaling development

- Extending Xen with BTS support
 - Debug crashes in Xen, and dom0
- Execution comparison tools
 - BTS traces to understand what went wrong
 - 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, dom0, domU
- Support performance tracing
 - Xentrace messages

What we didn't predict

- I/O delay goes up
 - Not sure if Linux has adequate low-latency user-level processing support
 - Maybe need an in-kernel interposition component
- Branch counters are fragile
 - Our code works on several server CPUs
 - Fails on a laptop with the CPU from the same model/family line

Conclusions

Practical replay analysis is feasible

- Performance overheads are reasonable
 - Realistic systems
 - Realistic workloads
 - Minor setup costs (just install Xen)
- Analysis engine is an amazing tool
 - And we're growing it
- We need your help to port it upstream
 - Porting effort
 - Shadow memory for PV guests
 - Support for HVM guests

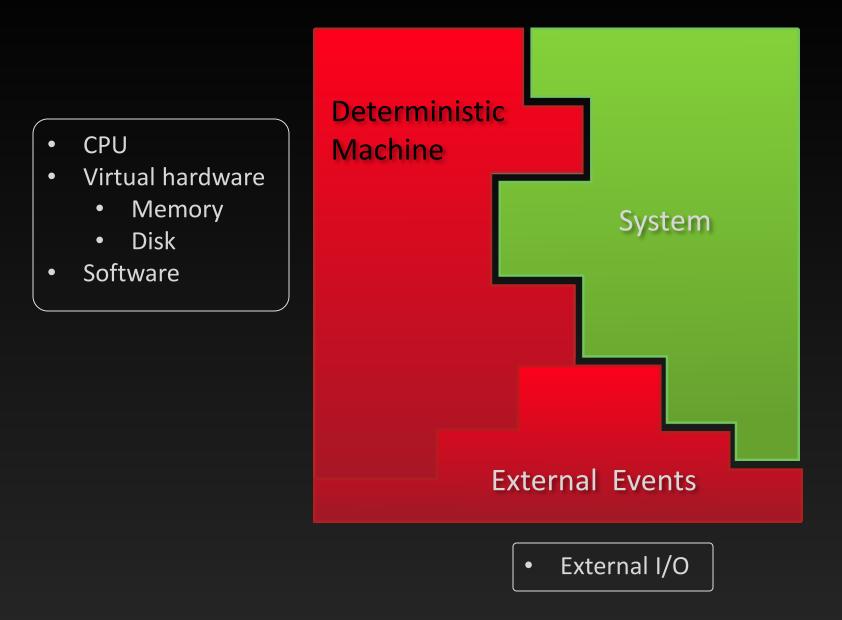
Thank you.

All code is GPLv2 and will be available soon (available on-request now).

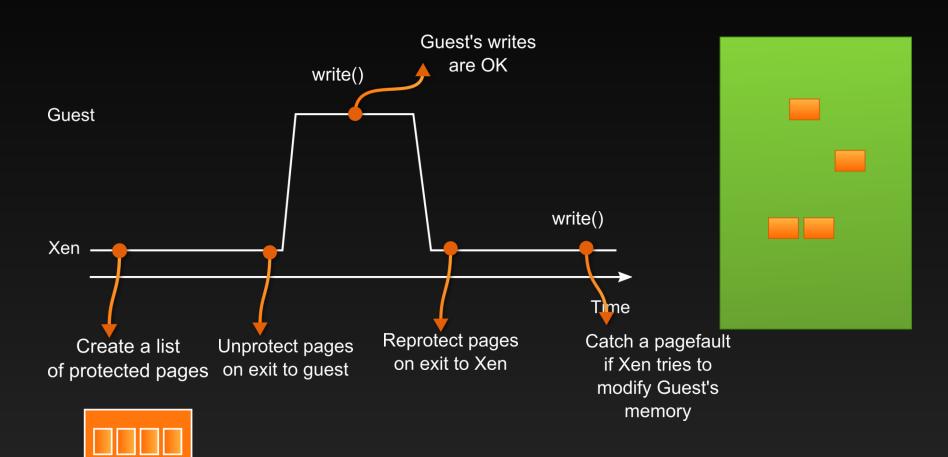
Questions: aburtsev@flux.utah.edu

Backup slides

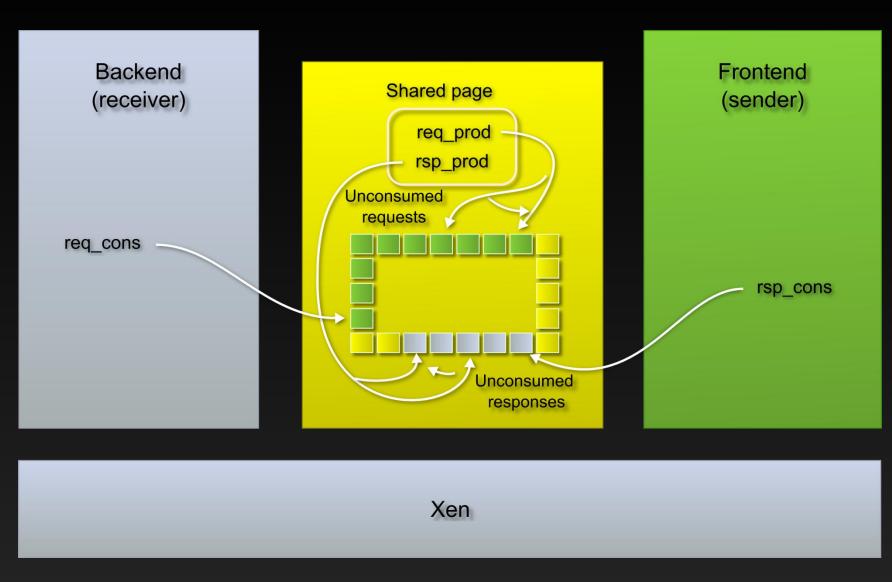
Execution environment



How do we find nondeterministic events?



Device interface



Reading a local variable

lval_skb = bsymbol_load(bsym_skb, flags);

skb_addr = *(unsigned long*) lval_skb->buf;

Controlled re-execution or replay

- Types of non-deterministic events
 - Synchronous
 - Hypercalls
 - Asynchronous
 - Interrupts
 - Best effort
 - Time updates
- Branch counters
 - The biggest problem of this solution