Toward Reproducible Malware Forensics

Brendan Dolan-Gavitt,
Columbia University
ACSAC MMF 2014

Reproducibility

- Basic ingredients:
    1. Describe your methods well
    2. (Optional, but highly recommended) Release your code
    3. Release your data
- (1), (2) examined in previous work (Collberg et al., 2014)
- But in the context of malware analysis, what does (3) mean?

Reproducibility Problems

- Software execution is ephemeral
  - Environment may change
  - Timings may change
  - Library versions, time of day, etc.
- Thus, *dynamic* analyses are hard to reproduce

Reproducible Malware Analyses

- Malware has short “shelf life”
  - C&C servers are quickly taken down
- Lack of access to reliable data sets makes research harder
  - Barriers to entry: need your own malware feed
  - Can’t tell if previous results are correct

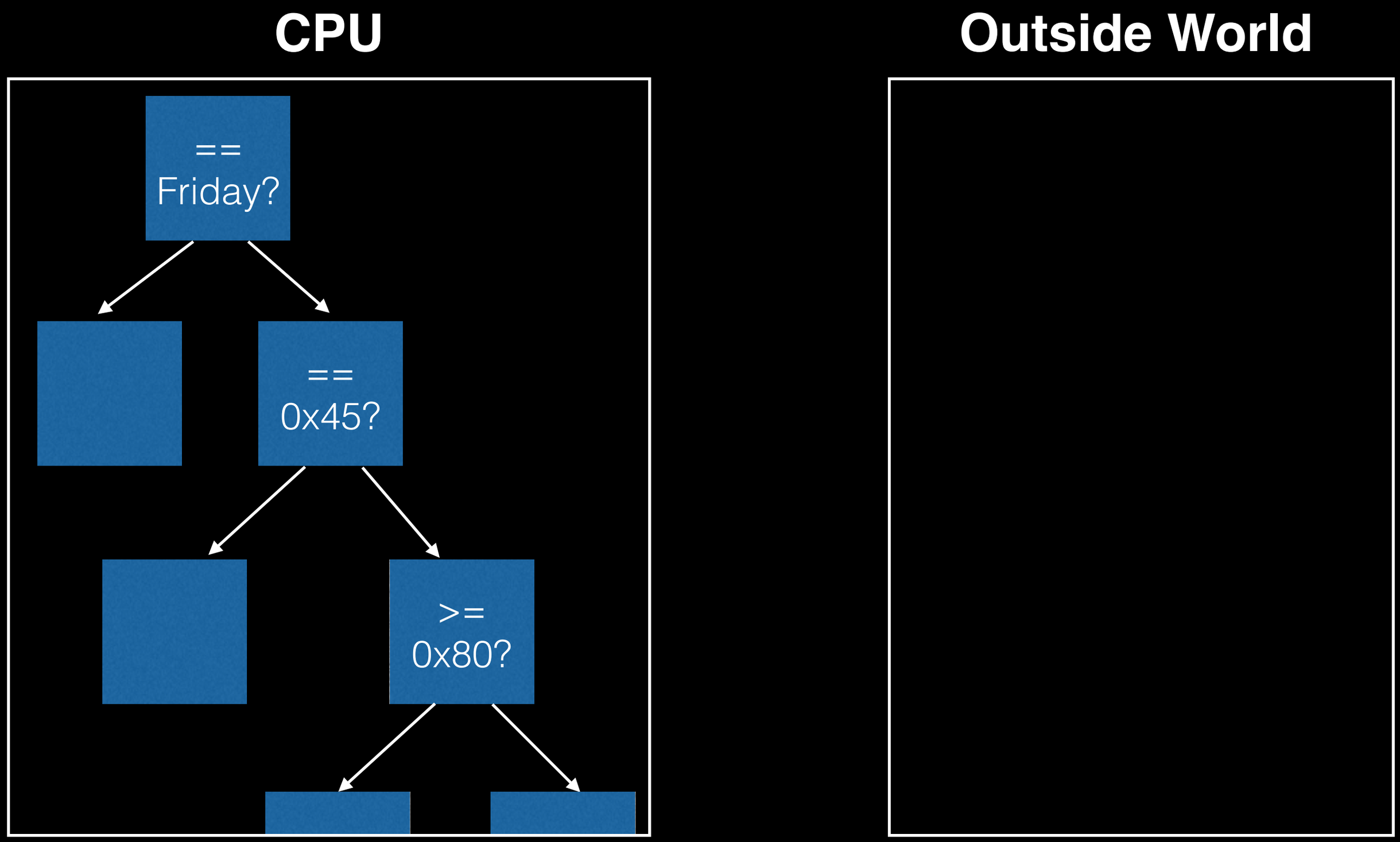
Previous Efforts

- Generally *artifact*-based or assume static analysis
- Malware sample repositories
  - VXShare, OpenMalware (née Offensive Computing), Contagio, …
- Malware artifact repositories
  - DHS Predict (PCAPs), Malwr (behavioral reports)

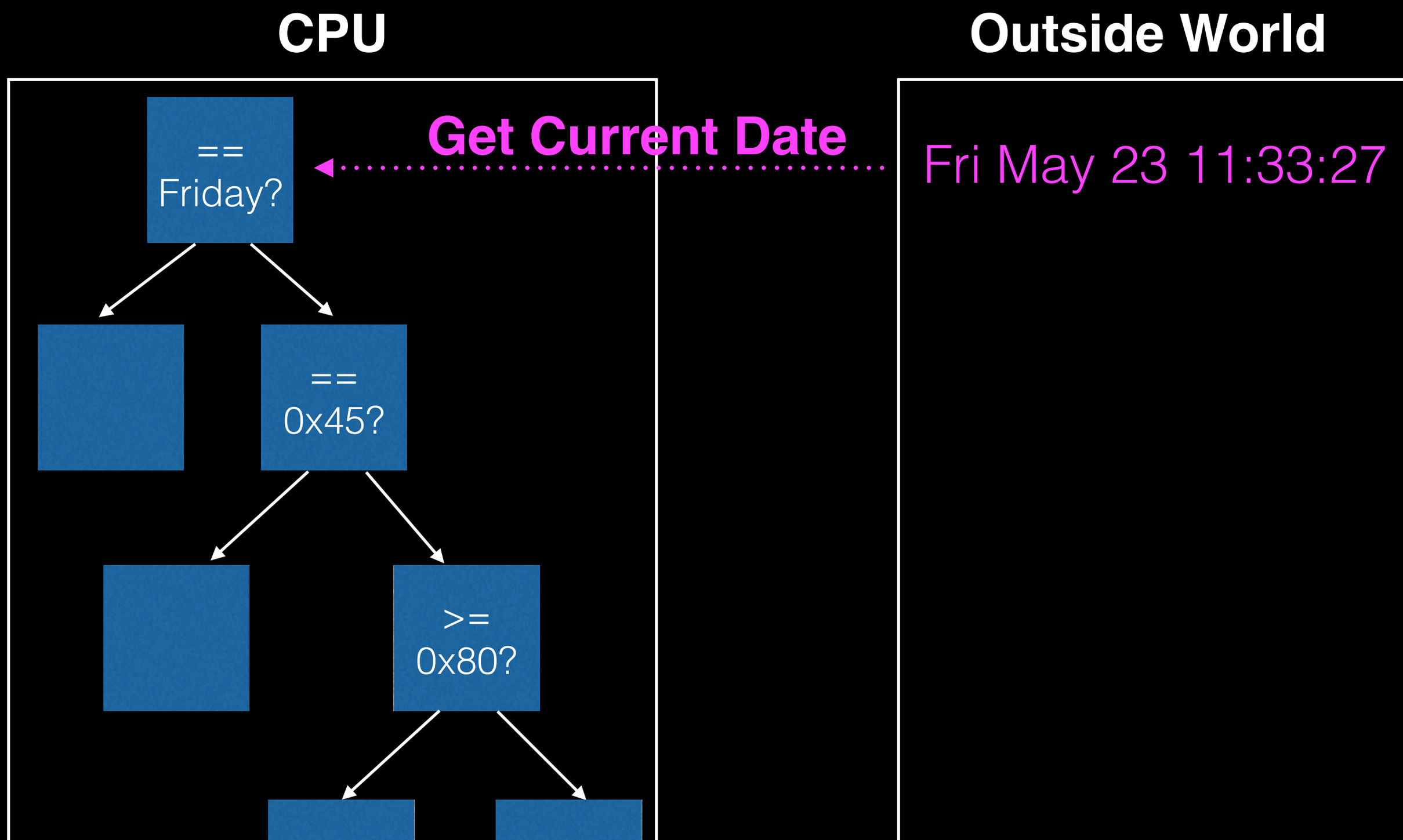
Idea: Shareable Record/Replay

- Full execution traces would solve this, but are enormous (GB/s)
- Instead, use record/replay: classic (30+ years old) technique for recording program executions
- Lots of academic literature on it: ReVirt, TTVM
- Main idea: record the *non-deterministic inputs*
- Until recently, no open source whole-system implementations

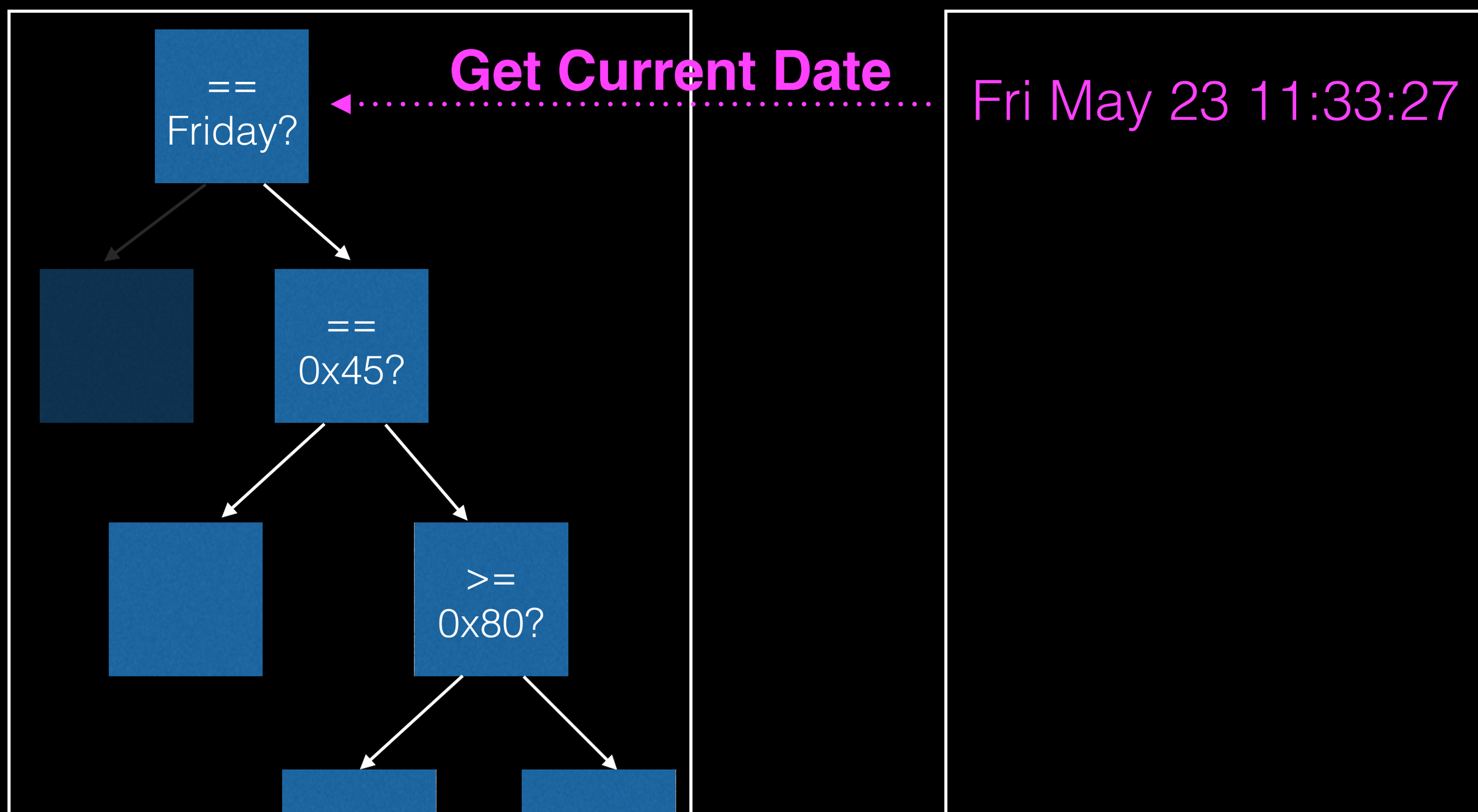
Record/Replay



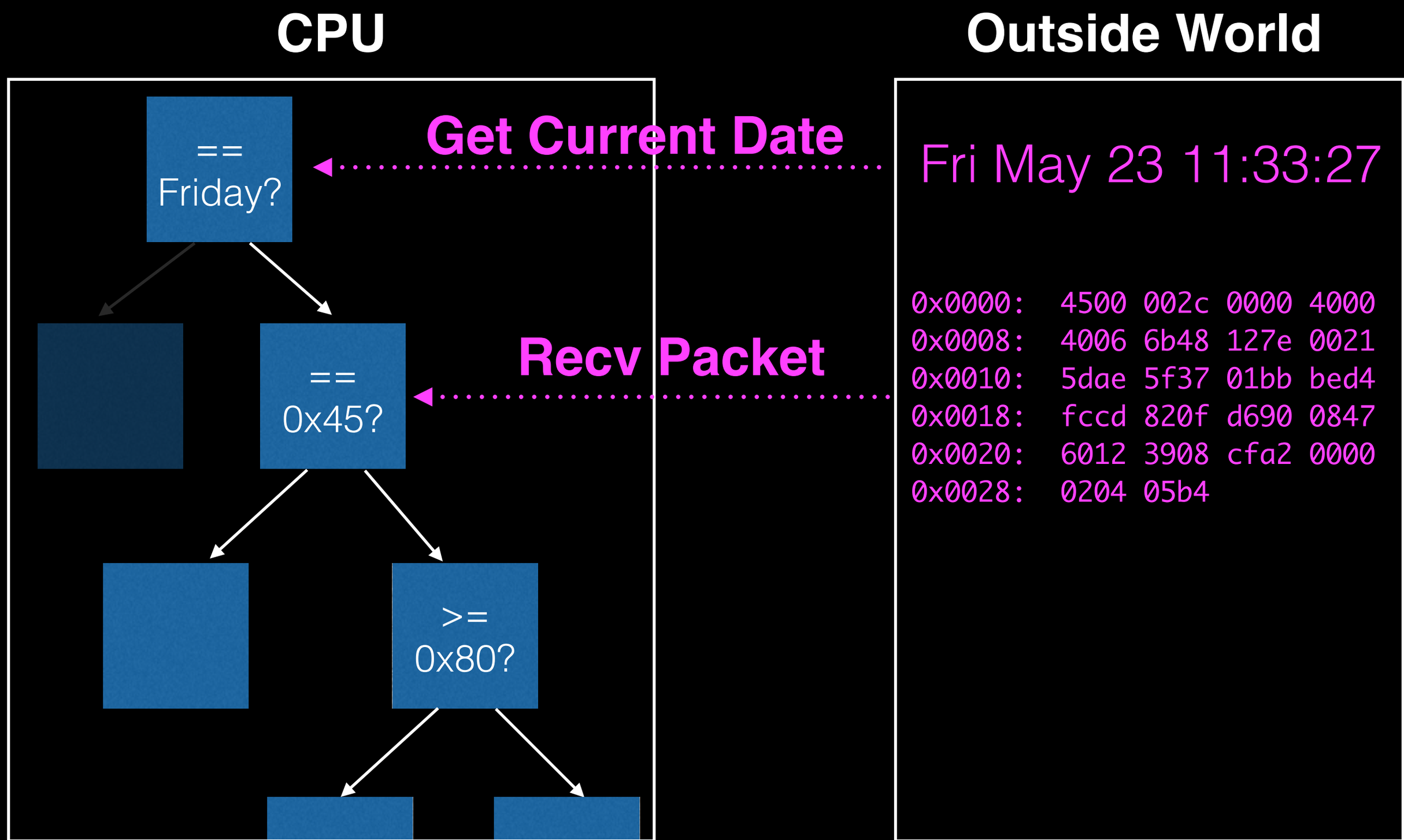
Record/Replay



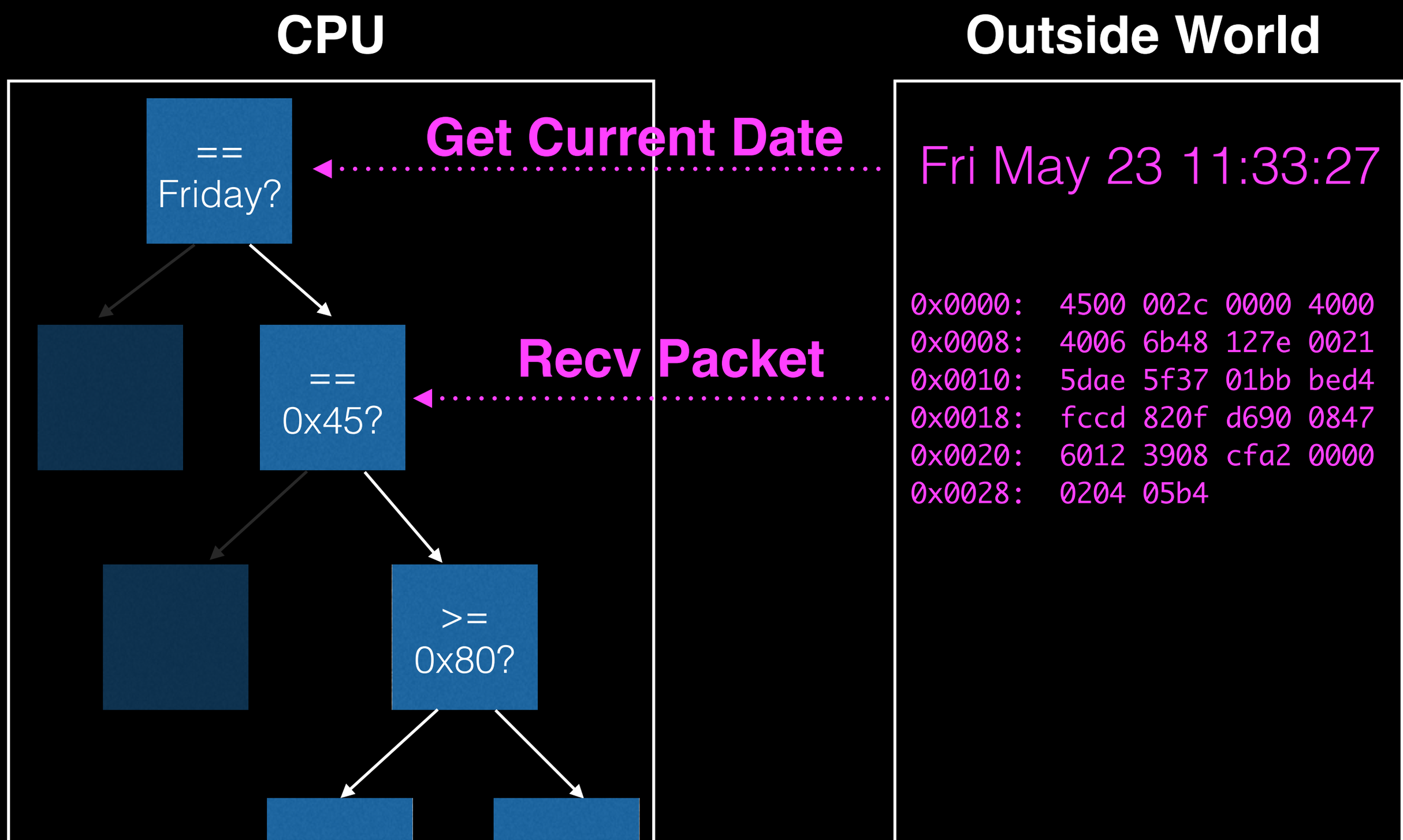
Record/Replay



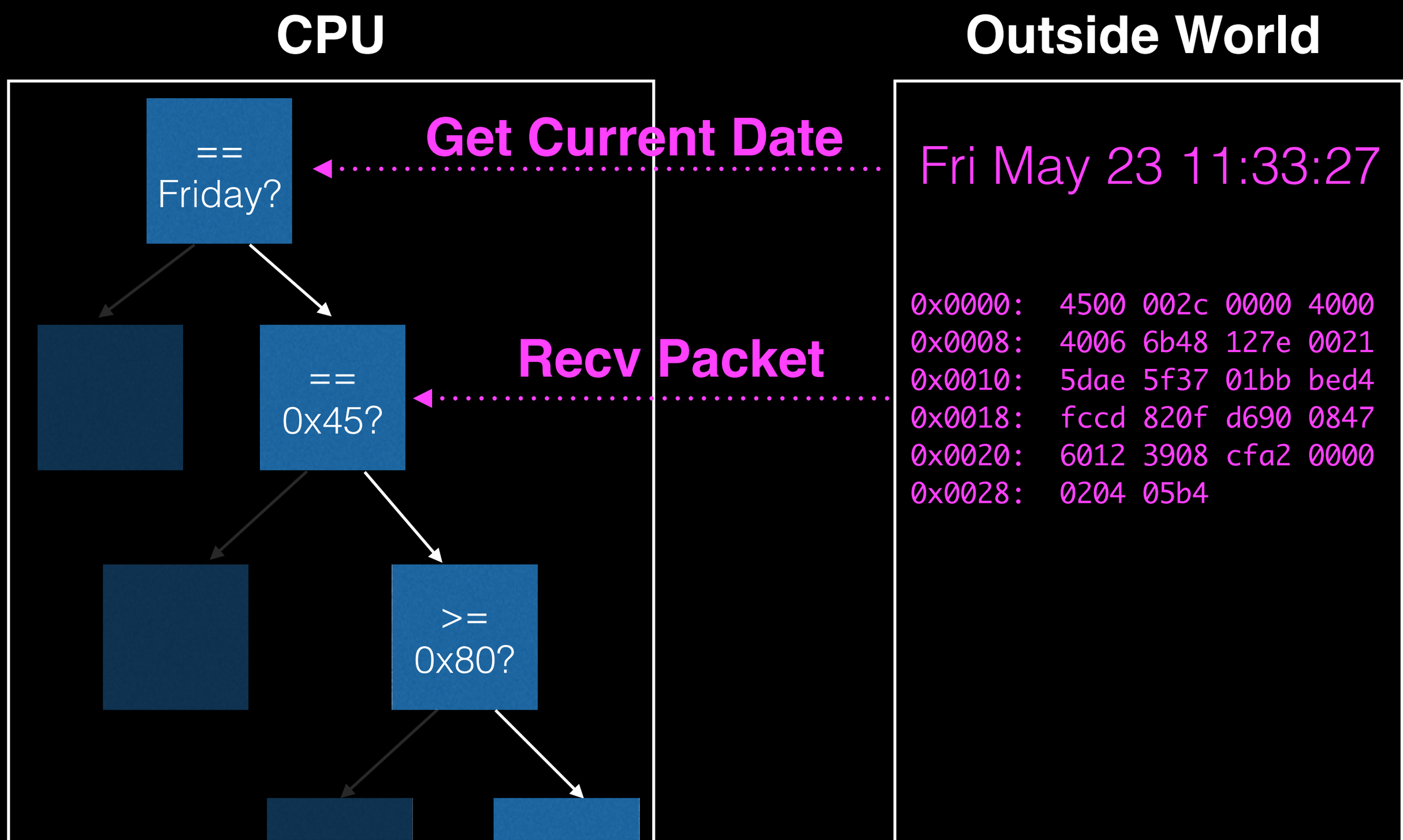
Record/Replay



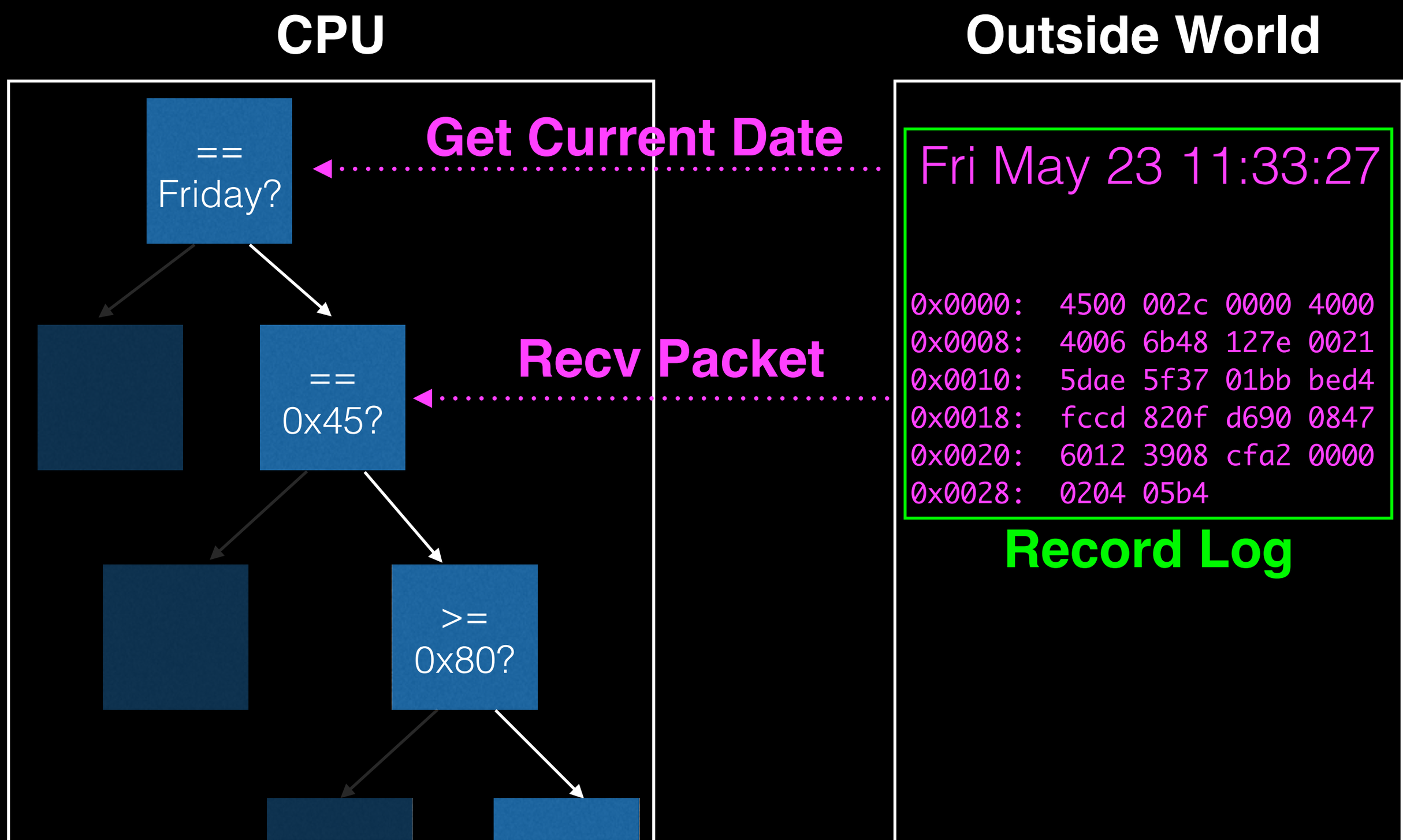
Record/Replay



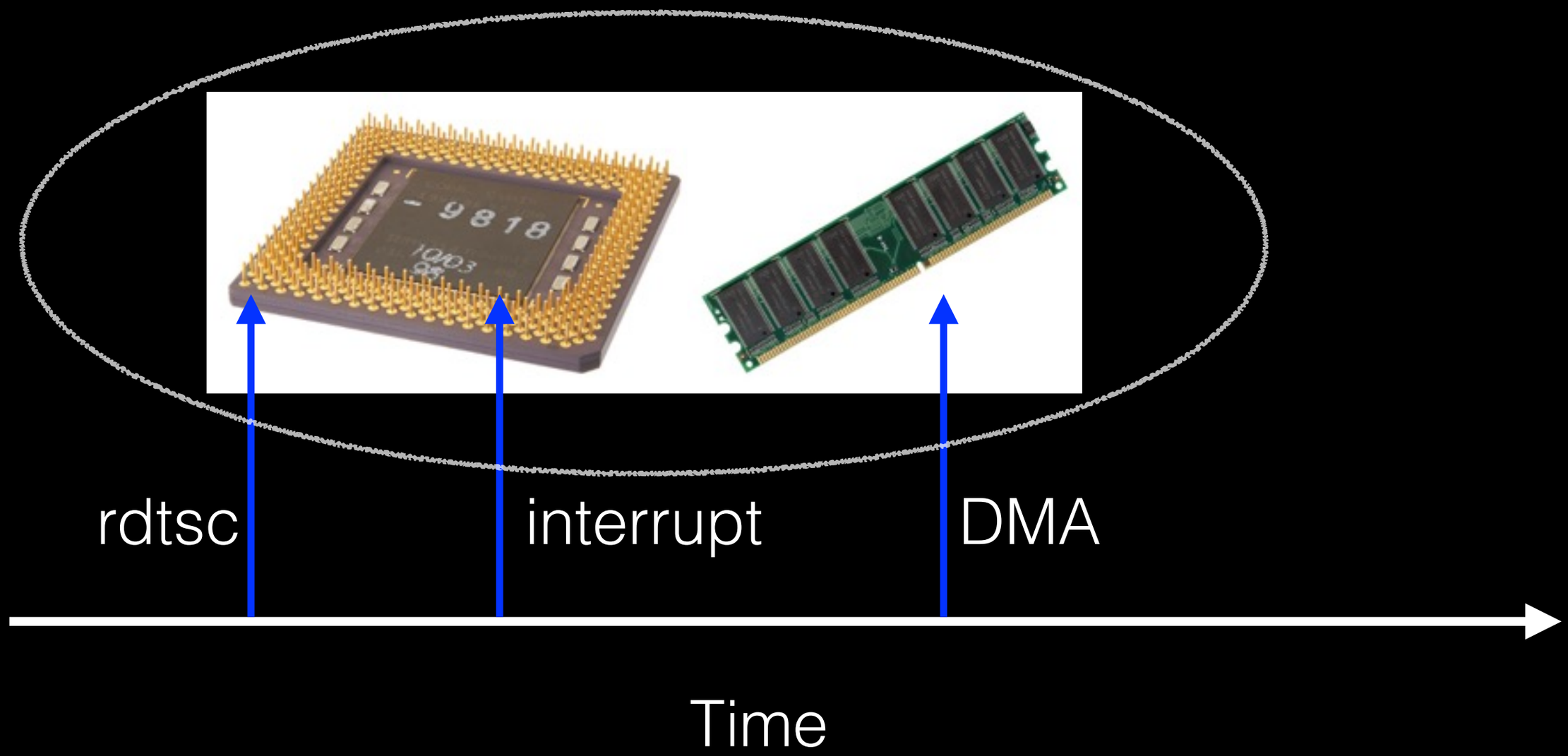
Record/Replay



Record/Replay



Record / Replay



PANDA

- Based on QEMU 1.0.1
- **Deterministic record/replay**
- Translation to LLVM for all QEMU architectures (extended from S2E code)
- Android (ARM) emulation support
- Plugin architecture – easy to extend to new analyses

Log Size

Replay	Instructions	Log Size	Instr/Byte
freebsdboot	9.3 billion	533 MB	17
spotify	12 billion	229 MB	52
haikuurl	8.6 billion	119 MB	72
carberp1	9.1 billion	43 MB	212
win7iessl	8.6 billion	9.4 MB	915
Starcraft	60 million	1.8 MB	33

PANDA Share - Share PAN

www.rrshare.org

Logged in as moyix
Logout

PANDA SHARE

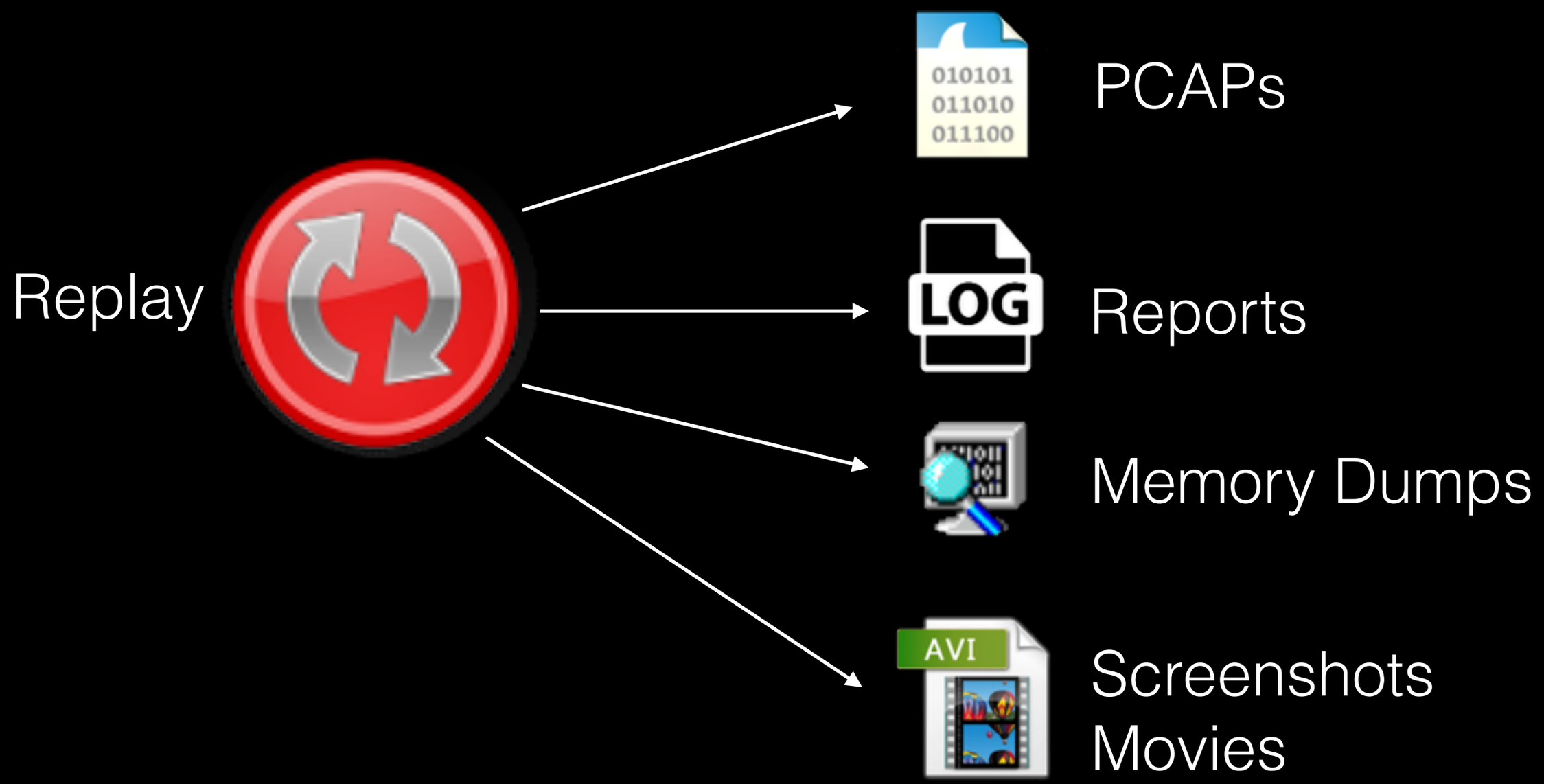
[Home] [About]

This site stores recordings made with the PANDA dynamic analysis platform. To find out more about PANDA's record/replay features, you can peruse the documentation. After downloading, the `.rr` files can be extracted using scripts/rrunpack.py in the PANDA distribution.

Upload a new record/replay log

Name	Summary	Download	Size	Instructions
cve-2012-4792-exploit	Exploitation of cve-2012-4792	rrlogs/cve-2012-4792-exploit.rr	130.1 MB	968.8 million
cve-2012-4792-crash	Crashing instance of cve-2012-4792	rrlogs/cve-2012-4792-crash.rr	129.9 MB	608.8 million
cve-2011-1255-exploit	Exploitation of cve-2011-1255	rrlogs/cve-2011-1255-exploit.rr	126.6 MB	2.1 billion
cve-2011-1255-crash	Crashing instance of cve-2011-1255	rrlogs/cve-2011-1255-crash.rr	127.1 MB	1.4 billion
cve-2014-1776-crash	Crashing instance of cve-2014-1776	rrlogs/cve-2014-1776-crash.rr	155.9 MB	1.2 billion
dia2dump	Parsing a PDB with dia2dump	rrlogs/dia2dump.rr	190.8 MB	5.4 billion
line2	Sending an IM using LINE for Android	rrlogs/line2.rr	64.6 MB	10.4 billion
win7_64bit_install_STOP_D1	Failure during boot to install CD of Win7 64bit. DRIVER_IRQL_NOT_LESS_OR_EQUAL	rrlogs/win7_64_install_fail.rr	203.3 MB	5.3 billion
carberp2	Running custom RU_Az build of the Carberp malware	rrlogs/carberp2.rr	91.9 MB	2.9 billion
	Running custom Full build of the Carberp			

Replay Subsumes Other Artifacts



Recycle Bin

Adobe Reader XI

Administrator: Command Prompt

```
C:\Windows\system32>_
```

Windows 7
Build 7601
This copy of Windows is not genuine

8:46 PM
11/24/2014

Recycle Bin

Adobe Reader XI

Administrator: Command Prompt

```
C:\Windows\system32>_
```

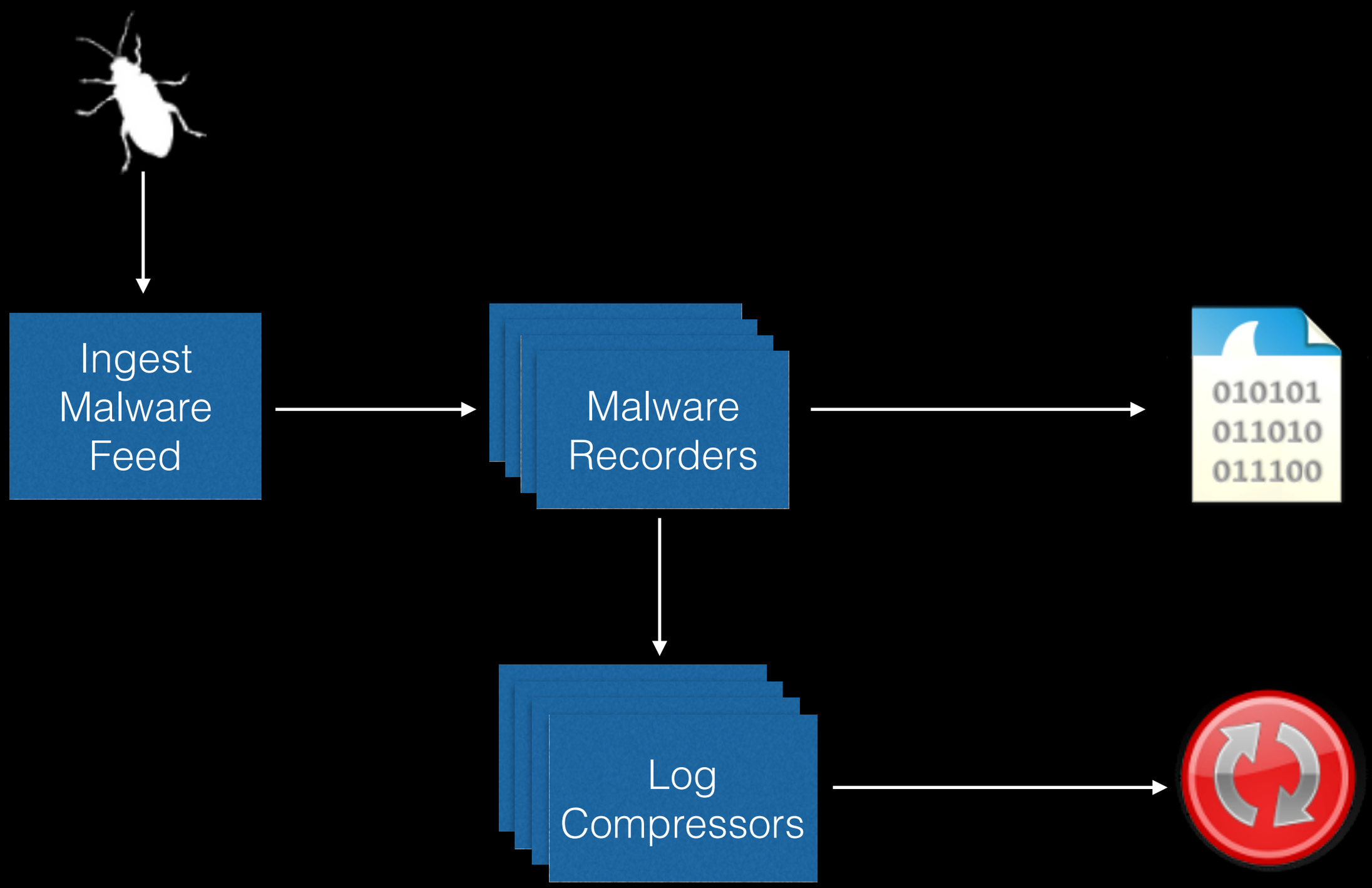
Windows 7
Build 7601
This copy of Windows is not genuine

8:46 PM
11/24/2014

MalRec: A Malware Recording Platform

- Based on PANDA dynamic analysis platform
- Simple agentless setup:
  - Malware loaded via CD image
  - Started by sending keystrokes to VM
  - No in-guest monitoring utilities (reports can be generated from replays)

Malware Pipeline



Implementation Details

- Samples fetched once per day at 22:30 UTC, random subset of 100 chosen
- `inotifywait` monitors incoming directory and passes off samples to GNU `parallel` & PANDA
- PANDA runs using -record_from and base Win7 32- or 64-bit base QCOW2
- Resulting logs are compressed with xz through another inotify/parallel queue

panda.gtisc.gatech.edu/malrec/

UUID	Filename	MD5	PCAP	RR Log	Added
4fc89505-75a0-4734-ac6d-1ebbdca28caa	005b80688b590435b7aab13342a00c6e.exe	005b80688b590435b7aab13342a00c6e	pcap	rrlog	2014-12-08 01:32:51.913522107 +0000
a64339ce-5fcb-415e-99f4-aa639c635805	02b955cf0d29e46502cb5dafd4244082.exe	02b955cf0d29e46502cb5dafd4244082	pcap	rrlog	2014-12-08 01:36:18.581528091 +0000
9a5cfaee-a478-444f-8dca-7f401f8f0df5	00b68dc33cd0a7122ffc8f1a237528c7.exe	00b68dc33cd0a7122ffc8f1a237528c7	pcap	rrlog	2014-12-08 01:32:58.649522302 +0000
92b72e3a-917c-4792-91aa-1d9950739d99	005de27b207285e70dea705feff8a4e7.exe	005de27b207285e70dea705feff8a4e7	pcap	rrlog	2014-12-08 01:33:11.061522661 +0000
e2152d26-73ff-4953-907d-8d6e9e32a4f3	03627679800f9540633a0a338e2d1930.exe	03627679800f9540633a0a338e2d1930	pcap	rrlog	2014-12-08 01:36:51.157529034 +0000
bc2581aa-85e8-4012-9e27-c728a00f3ff8	02b9a077e3c373089f0624a8bb66ec8d.exe	02b9a077e3c373089f0624a8bb66ec8d	pcap	rrlog	2014-12-08 01:36:04.829527693 +0000
3be52156-4f93-4a37-9af1-d1d45b526825	03d33743572fa24494582f24137e0d89.exe	03d33743572fa24494582f24137e0d89	pcap	rrlog	2014-12-08 01:32:47.105521968 +0000
f8b6036a-40d8-486b-af0c-8ec2840960f4	03d78a0f036ea665b8147a584584b179.exe	03d78a0f036ea665b8147a584584b179	pcap	rrlog	2014-12-08 01:36:24.365528258 +0000
f2d1662f-1079-4f45-b542-8b1cf8fdb1a9	079e0f2a6d817d8c88b1587f352d7cd0.exe	079e0f2a6d817d8c88b1587f352d7cd0	pcap	rrlog	2014-12-08 01:40:47.225535869 +0000
5d5eb4f6-13b0-44ed-bfd6-73b5aa0d284f	0995d976f26730007596d14fcce219a0.exe	0995d976f26730007596d14fcce219a0	pcap	rrlog	2014-12-08 01:40:24.841535221 +0000
a6d2a1e0-027c-4f80-90c6-9e9f84de53da	0c5fd363447293ac308e8079d532192c.exe	0c5fd363447293ac308e8079d532192c	pcap	rrlog	2014-12-08 01:40:09.885534788 +0000
d4ec17b9-90ec-4e96-b40b-f6e77f5ca1a7	0e1d93833d3909e454b79c9ccf82c698.exe	0e1d93833d3909e454b79c9ccf82c698	pcap	rrlog	2014-12-08 01:40:10.293534800 +0000
781b95ff-943f-4590-877e-442d31991320	0f3f08e54ac62879b8ac4873e4be58e9.exe	0f3f08e54ac62879b8ac4873e4be58e9	pcap	rrlog	2014-12-08 01:43:53.813541271 +0000
0c413017-1c47-4d48-b90e-5d21e5407b52	101357b66a53eb86cab6c69fc48df3b7.exe	101357b66a53eb86cab6c69fc48df3b7	pcap	rrlog	2014-12-08 01:42:42.201539198 +0000
60a022d2-2287-4814-8d0d-676e215c0db1	10146d57a77bd3008e7f789b2a1b2540.exe	10146d57a77bd3008e7f789b2a1b2540	pcap	rrlog	2014-12-08 01:42:54.657539558 +0000
8edbd0f0-9d0f-41d9-9148-bc92966e949b	12b5501c2f30e8c3b7a8475da1c8e05e.exe	12b5501c2f30e8c3b7a8475da1c8e05e	pcap	rrlog	2014-12-08 01:57:43.441565292 +0000
537a5f48-7233-4996-af8e-20e3df1e99aa	11b64c44a79fc463d1c46c9faf1856ca.exe	11b64c44a79fc463d1c46c9faf1856ca	pcap	rrlog	2014-12-08 01:57:39.601565180 +0000
f481da2e-5ca1-4e60-a7d9-45a3a410f758	11225eec69d383c79fb6d4bff180ca7d.exe	11225eec69d383c79fb6d4bff180ca7d	pcap	rrlog	2014-12-08 01:57:41.937565248 +0000
f2298ba9-af24-473b-b14e-b564445741c8	17af4487d844314a20f03c866d3d5fa2.exe	17af4487d844314a20f03c866d3d5fa2	pcap	rrlog	2014-12-08 01:57:43.593565296 +0000
f220daf4-eaff-4626-b935-6938e5fd5c2f	257db161cbcf9d820b00c51b6d7d18e7.exe	257db161cbcf9d820b00c51b6d7d18e7	pcap	rrlog	2014-12-08 02:03:16.933574947 +0000
b437845a-6c4e-48c2-b1cf-db8e18e369df	1e888f5b607899b50c09f1840b474d0c.exe	1e888f5b607899b50c09f1840b474d0c	pcap	rrlog	2014-12-08 02:04:05.617576357 +0000
9f5b9ff9-957f-4b4d-8c50-6f028ab134e2	1c012c325a06e52b1e56b1a3420620e2.exe	1c012c325a06e52b1e56b1a3420620e2	pcap	rrlog	2014-12-08 02:03:25.617575199 +0000
0d8cf2c9-b9c0-468b-8b55-9a9c2f7b0459	267c351d05b28db0c06620536bf4f010.exe	267c351d05b28db0c06620536bf4f010	pcap	rrlog	2014-12-08 02:03:42.361575683 +0000
8cba72d5-9f8d-446d-a9fe-7abf85d025fc	26e7f238b29cdc9c9ca06b35332f0c77.exe	26e7f238b29cdc9c9ca06b35332f0c77	pcap	rrlog	2014-12-08 02:08:39.733584293 +0000
464d62fe-20e9-43a7-afb1-ae730e571163	29cc460c9fa5c6b7edea77eaf91102c9.exe	29cc460c9fa5c6b7edea77eaf91102c9	pcap	rrlog	2014-12-08 02:08:59.489584865 +0000
813e63fc-43aa-498a-8af2-d8088384b874	289510340cc1396f995bf20ee4ea9bb3.exe	289510340cc1396f995bf20ee4ea9bb3	pcap	rrlog	2014-12-08 02:09:10.321585179 +0000
ce28db56-a5d3-4a28-ba69-3f603192e3ce	29dc3212b5fae469ecffa8ed1a1a1599.exe	29dc3212b5fae469ecffa8ed1a1a1599	pcap	rrlog	2014-12-08 02:09:11.157585203 +0000
974dbfac-4017-441e-8471-f84c81c7a818	2b4c8a076d21ccaf82e6e60b05d9f033.exe	2b4c8a076d21ccaf82e6e60b05d9f033	pcap	rrlog	2014-12-08 02:14:00.745593588 +0000
7c8801fc-c29f-49c5-8412-dce75dea3fa0	3b5c8f00989260c51395cd0d09aa0cb1.exe	3b5c8f00989260c51395cd0d09aa0cb1	pcap	rrlog	2014-12-08 02:14:06.501593754 +0000
e5f6f3d3-29e4-42fd-9011-522054fee9f3	2db49478ce69cb1beaa3e96471cdf4e2.exe	2db49478ce69cb1beaa3e96471cdf4e2	pcap	rrlog	2014-12-08 02:14:05.069593713 +0000
9fc52909-6fa1-468f-b5dc-280b7d0c2e17	3ba61a3efa0227bd4d7e0a3e2d6e415c.exe	3ba61a3efa0227bd4d7e0a3e2d6e415c	pcap	rrlog	2014-12-08 02:14:19.137594120 +0000
5bc23607-cc4c-468c-b25c-3351920bb6ba	3d3f5e93b5386db5fdc8e637a5ed0480.exe	3d3f5e93b5386db5fdc8e637a5ed0480	pcap	rrlog	2014-12-08 02:19:19.849602827 +0000

Limitations

- Analysis time is fixed & no interaction is done
  - In particular, only one path through malware
- PANDA is based on QEMU 1.0.1 & non-virtualized – very detectable
- *Lock-in*: replay logs can currently only be processed by PANDA, so initial analyses must be done in PANDA as well

Future Work

- Scaling up
  - Currently, limited by disk space (~20GB/day)
  - We get ~2000 samples/day, only record 100
- Add automated reports
  - Currently some basic support in PANDA, e.g. http://laredo-13.mit.edu/~brendan/opcleaver/reports/0c3e4035-c9ab-47bc-b245-35c80ceafe5e_proclog.txt
  - Movies of executions

Future Work

- Malware “mind reading”
  - Record all memory reads/writes and look for printable strings
  - Save all printable strings
  - Index and use information retrieval (e.g., Lucene or Terrier) to build search engine for malware memory accesses

Obtaining the Recordings

- http://panda.gtisc.gatech.edu/malrec/
- If you want easier bulk access, contact me at brendan@cs.columbia.edu and we can arrange something like an rsync transfer.

Questions
Discussion