TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Institute for System Architecture, Operating Systems Group

Paper Reading Group

Winter 2014

Dresden, 2014-10-15

Administrativia

- Every Wednesday, 11:10 AM, INF/3105
- Presentation + discussion of one paper
- Staff papers voted on
- Mailing List (see website)
- Usually: Pizza

Rules – Johns Hopkins University

- Pick one paper
  - **Explore** field of research by following related work
  - **Present** research field in 75 minutes talk
  - **Write** 8 page survey paper
- Pick own topic
  - Write a paper suitable for **workshop submission**

Rules – TUD

- One paper presentation per student
- Pick a paper related to systems research
  - Suggestions on the website
- Prepare ~15 min presentation
  - In English
  - Show that you understood the paper
  - Extra knowledge (related work) may be helpful
  - Prepare questions/issues for discussion

Rules – TUD (ctd.)

- For papers you do not present:
  - Write a paper summary
    - Explain what you understood
    - Raise questions / issues
    - Mention things you liked / disliked
  - Send to Björn Döbel (doebel@tudos.org)
    - Deadline: midnight before the meeting (Tuesday 23:59:59)

TECHNISCHE UNIVERSITÄT DRESDEN

Faculty of Computer Science Institute for System Architecture, Operating Systems Group

Torturing Databases for Fun and Profit

Mai Zheng, Joseph Tucek, Dachuan Huang, Feng Qin, Mark Lillibridge, Elizabeth S. Yang, Bill W Zhao, Shashank Singh

The Ohio State University and HP Labs

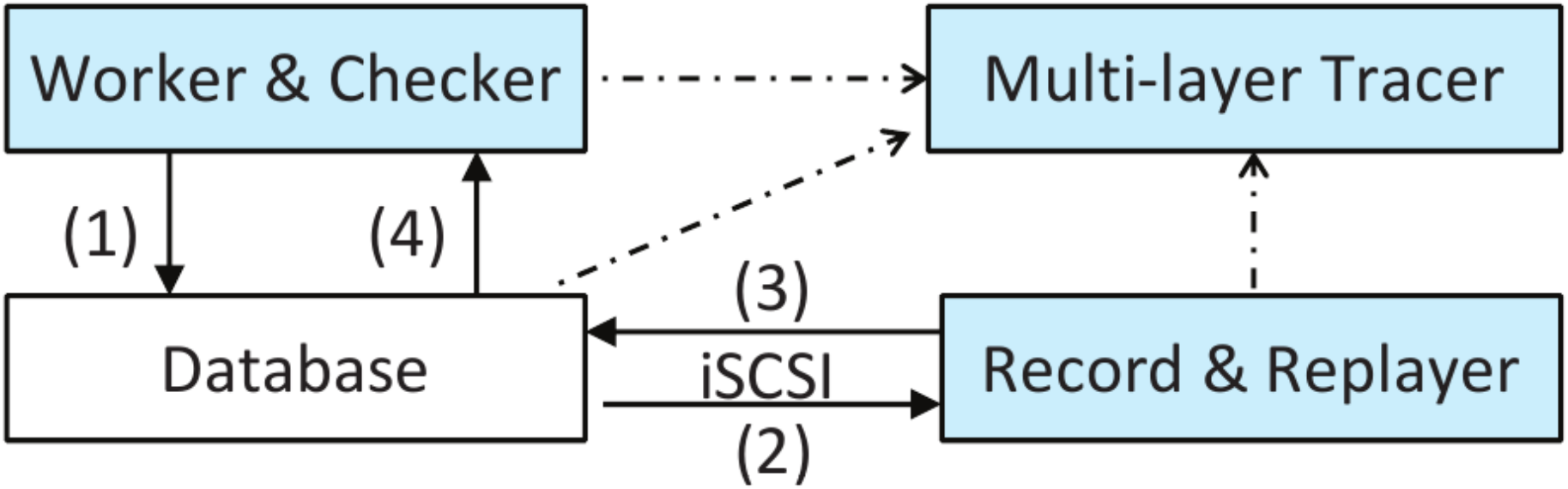
Dresden, 2014-10-15

ACID

- ATOMICITY
  - all or nothing
- CONSISTENCY
  - invariants hold between transactions
- ISOLATION
  - no intermediate states visible
- DURABILITY
  - committed is committed

Methodology

- Fault Model: Clean Power Faults



Workloads

- W1: Single thread, single transaction

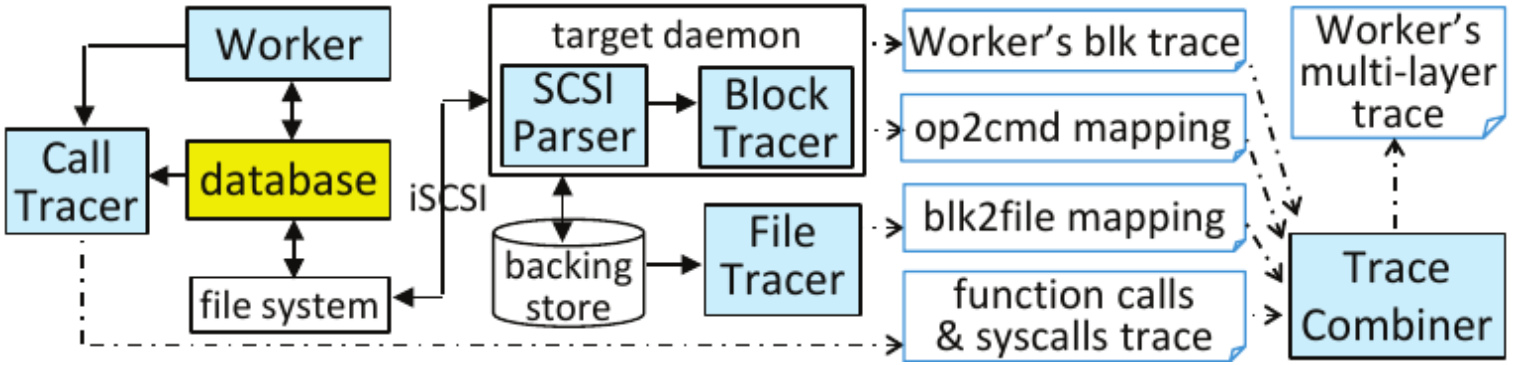
```
Begin Transaction
  for i = 1 to txn_size do
    key = "k-" + str(i)
    value = "v-" + str(i)
    put(key, value)
  end
  before_commit = get_timestamp()
Commit Transaction
after_commit = get_timestamp()
```

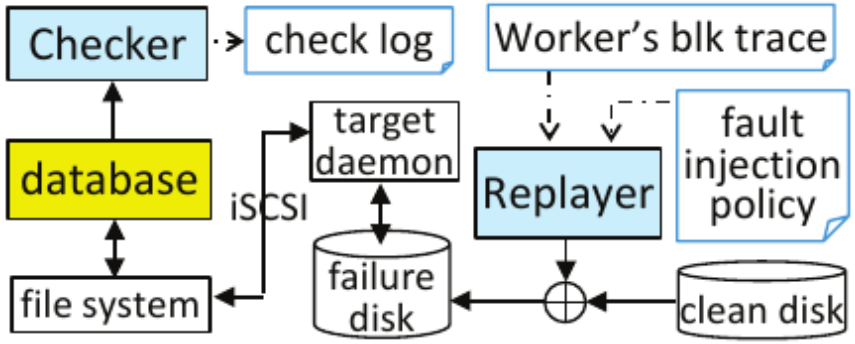
- Check atomicity, consistency and durability

Worloads (ctd.)

- W2: Multithreaded version of W1
  - no concurrent transactions overlapping
  - stresses concurrency handling
- W3: Single-threaded multi-row consistency
  - concurrent non-overlapping bank txns
- W4: Multi-threaded overlapping

Record / Replay





Fault Injection Policy

- Exhaustive (can take several months)
- Pattern based:

op#	LBA	file
...	...	...
35	***1038***	***x.db***
36	2347	x.db
37	2351	x.db
...	...	...
49	1038	x.db

(a) P_{rep}

op#	LBA	file
61	3142	x.db
62	3146	x.db
63	***2081***	***x.db***
64	***5191***	***x.db***
65	1025	x.db
66	1029	x.db

(b) P_{jump}

op#	LBA	cmd#
152	1070	42
153	***1106***	***43***
154	***1110***	***43***
155	***1114***	***43***
156	1118	43
157	1765	44

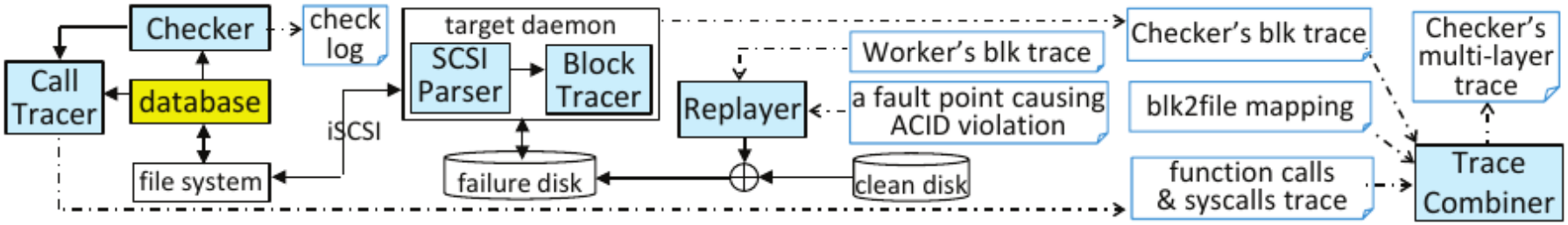
(c) P_{head}

op#	LBA	file
245	5545	x.db
246	5646	x.db
247	***5545***	***x.db***
248	***8351***	***fs-j***
249	***8352***	***fs-j***
250	8356	fs-j

(d) P_{tran}

Unintended update to mmap'ed blocks

Diagnosis Replay



TECHNISCHE UNIVERSITÄT DRESDEN

Results (Exhaustive)

DB	FS	W-1	W-2	W-3	W-4.1	W-4.2	W-4.3	A	C	I	D
TokyoCabinet	ext3	D	D	D	A C D	A C D	A C D	0.15	0.14	0	16.05
	XFS	—	D	D	A C D	D	A C D	<0.01	0.01	0	4.38
MariaDB	ext3	D	D	D	D	D	D	0	0	0	1.36
	XFS	D	D	D	D	D	D	0	0	0	0.49
LightningDB	ext3	—	—	—	—	—	D	0	0	0	0.05
	XFS	—	—	—	—	—	—	0	0	0	0
SQLite	ext3	D	D	—	D	D	D	0	0	0	19.15
	XFS	—	—	D	D	D	D	0	0	0	10.60
KVS-A	ext3	—	—	Hang*	—	—	—	0	0	0	0
	XFS	—	—	—	—	—	—	0	0	0	0
SQL-A	ext3	D	D	D	D	D	D	0	0	0	3.31
	XFS	D	D	D	D	D	D	0	0	0	0.92
SQL-B	ext3	D	D	C D	C D	C D	C D	0	8.96	0	3.24
	XFS	C D	D	C D	C D	C D	C D	0	7.77	0	3.90
SQL-C	NTFS	D	D	D	D	D	D	0	0	0	8.08

TECHNISCHE UNIVERSITÄT DRESDEN

Results (Patterns)

DB	W-4.1		W-4.3	
	match?	top?	match?	top?
TokyoCabinet	Y	Y	Y*	Y
MariaDB	Y	Y	Y	Y
LightningDB	—	—	Y	Y
SQLite	Y	Y	Y	Y
KVS-A	—	—	—	—
SQL-A	Y	N	Y	N
SQL-B	Y	N	Y*	Y
SQL-C	Y	Y	Y	Y

DB	Exhaustive	Pattern	%
TokyoCabinet	12d 1h*	2d 0h	16.6%
MariaDB	3h 27m	3m 2s	1.5%
LightningDB	7h 56m	20m 44s	4.4%
SQLite	13m 12s	0m 42s	5.3%
KVS-A	5h 17m	5m 32s	1.7%
SQL-A	3h 33m	10m 37s	5.0%
SQL-B	71d 1h*	2d 9h	3.4%
SQL-C	3h 23m	2m 34s	5.1%
Average	—	—	5.4%

Their contributions

- Workloads to test ACID properties
- Cross-platform methodology to expose reliability issues under power fault
- Pattern based ranking algorithm
- Multi-layer tracing system
- Experimental results against 8 databases

Discussion

- This is cool!
- Missing a discussion WHY xfs is better
- Shouldn't they also have found FS bugs?
- Don't they only check one interleaving (that of their recording)?