[wiredtiger-users] question about the key of first ref when split-insert

徐华叶 <xuhuaye@bytedance.com>
Reply-To: wiredtiger-users@googlegroups.com
To: wiredtiger-users <wiredtiger-users@googlegroups.com>

Hi everyone,

Recently I read the source-code of split-insert in WiredTiger, and I have a question about split-insert: why should split-insert change the key of first-ref? I noticed the comment in the source code, the reason is suffix-compressed, but I can't understand and also can't find a case in which the key of the first ref is invalid.

```
if (type == WT_PAGE_ROW_LEAF) {
    /*
     * Copy the first key from the original page into first ref in
     * the new parent. Pages created in memory always have a
     * "smallest" insert list, so look there first.  If we don't
     * find one, get the first key from the disk image.
     *
     * We can't just use the key from the original ref: it may have
     * been suffix-compressed, and after the split the truncated key
     * may not be valid.
     */
    WT_ERR(__wt_scr_alloc(session, 0, &key));
    if ((ins = WT_SKIP_FIRST(WT_ROW_INSERT_SMALLEST(page))) != NULL) {
        key->data = WT_INSERT_KEY(ins);
        key->size = WT_INSERT_KEY_SIZE(ins);
    } else {
        WT_ASSERT(session, page->entries > 0);
        WT_ERR(__wt_row_leaf_key(session, page, &page->pg_row[0], key, true));
    }
    WT_ERR(__wt_row_ikey(session, 0, key->data, key->size, child));
    parent_incr += sizeof(WT_IKEY) + key->size;
    __wt_scr_free(session, &key);
} else
    child->ref_recno = ref->ref_recno;
```

Let's suppose the origin key is A, the new key of the first ref is B. Because B is the smallest key of the page of first ref, so A must be less than or equal than B. In this way, I think A is always valid for the new first ref.

I found some issues of this, and the commit.

https://jira.mongodb.org/browse/WT-1419

https://jira.mongodb.org/browse/WT-1422

https://jira.mongodb.org/browse/WT-1421

https://github.com/wiredtiger/wiredtiger/commit/7ed4cc580e93e3bb1fc599f3ce99caaa9540d038

```
1078  1079          * The first page in the split is the current page, but we still need to
1079  1080          * create a replacement WT_REF and make a copy of the key (the original
1080  1081          * WT_REF is set to split-status and eventually freed).
1081  1082          *
1082  1083          * The new reference is visible to readers once the split completes.
1083  1084          */
1084  1085         WT_ERR(__wt_calloc_def(session, 1, &split_ref[0]));
1085  1086         child = split_ref[0];
1086  1087         *child = *ref;
1087  1088         child->state = WT_REF_MEM;
1088      -         __wt_ref_key(page, ref, &key, &size);
1089      -         WT_ERR(__wt_row_ikey(session, 0, key, size, &child->key.ikey));
      1089  +
      1090  +         /*
      1091  +          * Build a key to appear in the parent for the original page.
      1092  +          *
      1093  +          * We can't just use the key from the original ref: it may have been
      1094  +          * suffix-compressed, and after the split the truncated key may not be
      1095  +          * valid.
      1096  +          */
      1097  +         WT_ERR(__wt_scr_alloc(session, 0, &key));
      1098  +         if (page->pg_row_entries > 0)
      1099  +                 WT_ERR(__wt_row_leaf_key(
      1100  +                     session, page, &page->pg_row_d[0], key, 0));
      1101  +         else {
      1102  +                 ins = WT_SKIP_FIRST(WT_ROW_INSERT_SMALLEST(page));
```

From the issue, I know the first version of split-insert would cause the cache-test fail. I think the fail reason is this line "__wt_ref_key(page, ref, &key, &size);" . If we change the line to "__wt_ref_key(ref->home, ref, &key, &size);" , the failed test will succeed. From the source code we know, we should pass the parent page to the function "__wt_ref_key", not the child page. Later, I tested this idea, and the failed cache-test succeeded in fact.

I hope I have expressed myself clearly. Could someone tell me if I am right? Or are there some case that must change the key of first ref when split-insert?

Thank you!

huaye