

# CEPH中的写优化系统

2016-11-1

易怀杰

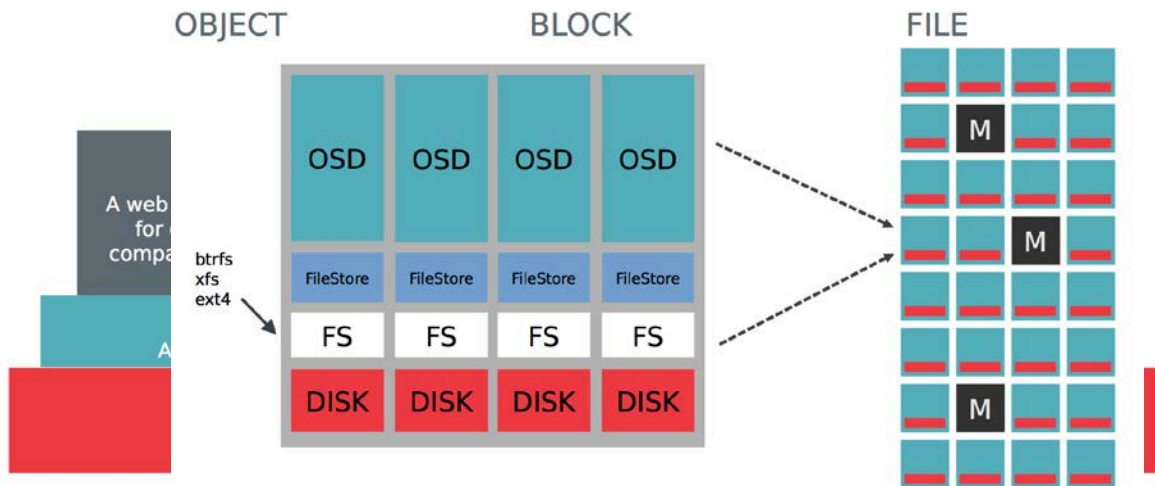
MG1533078

# 背景



## ➤ CEPH

1. 同一个集群
2. 所有的组件
3. 无单点故障
4. 运行在异构
5. 尽可能的自治。



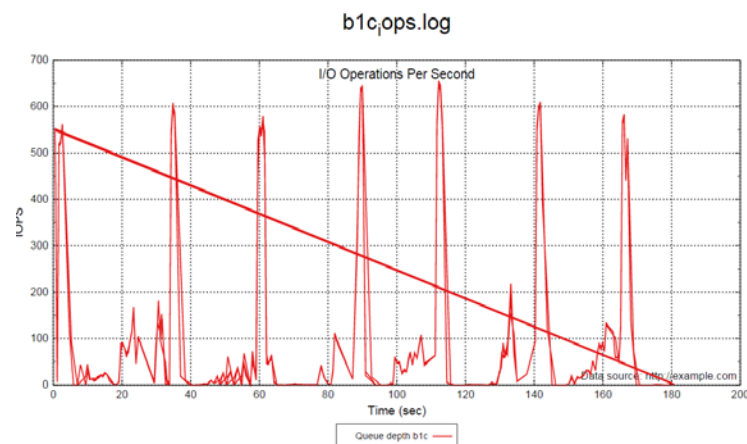
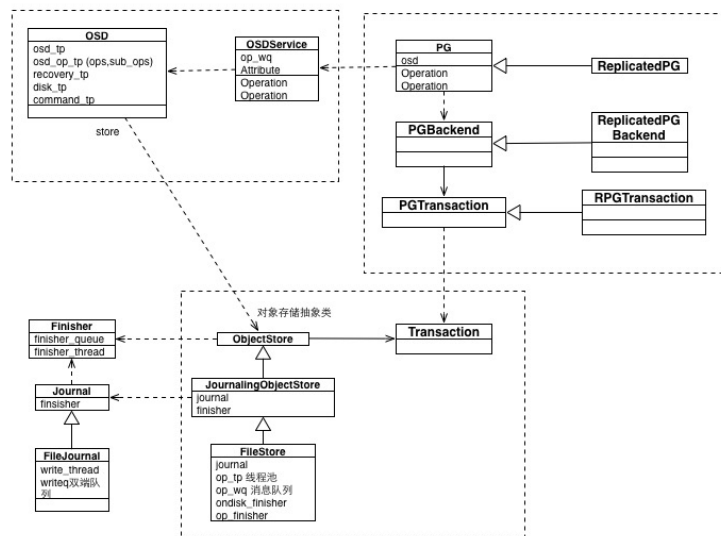
## ➤ POSIX文件系统的问题

1. 不合理的元数据管理——小文件的随机写性能极差。
2. 基于POSIX的WAL机制让磁盘I/O利用率下降了一半。

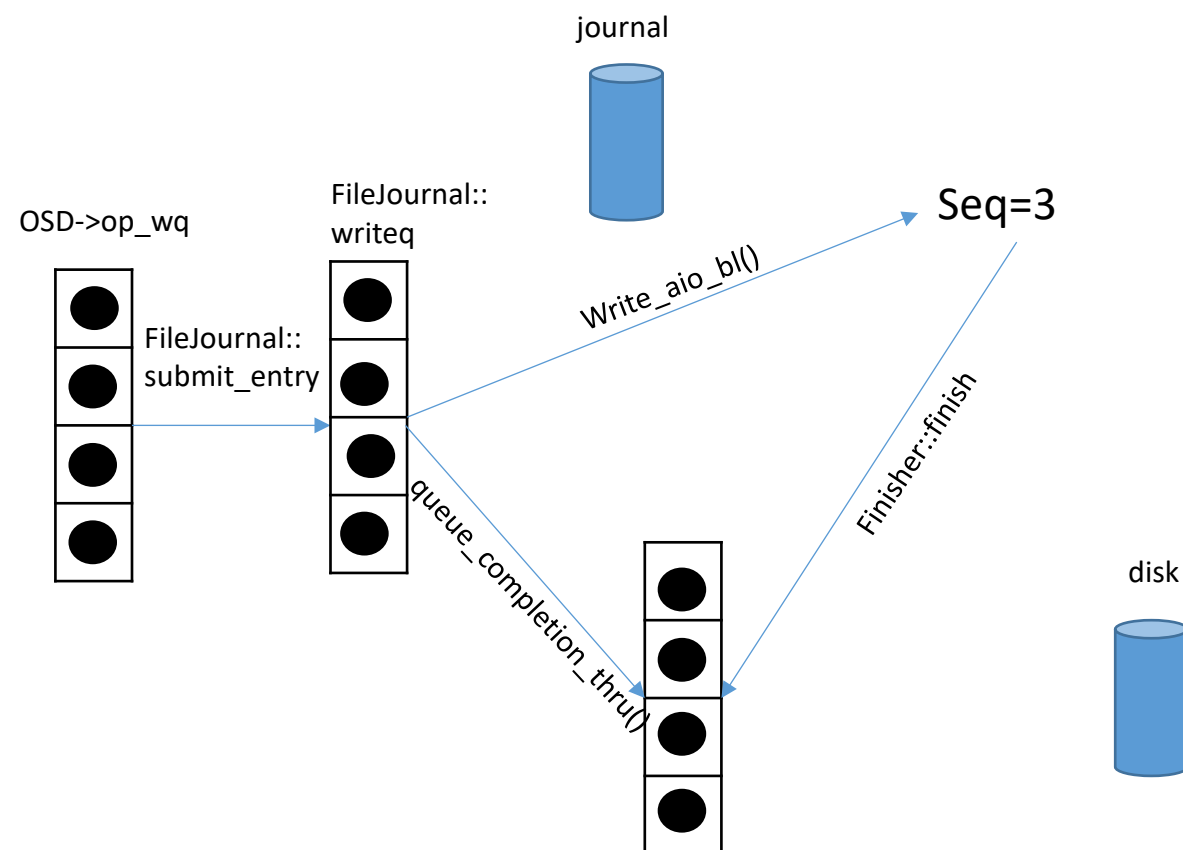
# 背景[2]

## ➤ FileStore的I/O抖动问题

- 3个OSD
- RBD+fio
- SSD做journal



# I/O抖动的原因



## ➤ 原因

1. WAL的二次写
2. journal落盘的机制不对

# 解决方案

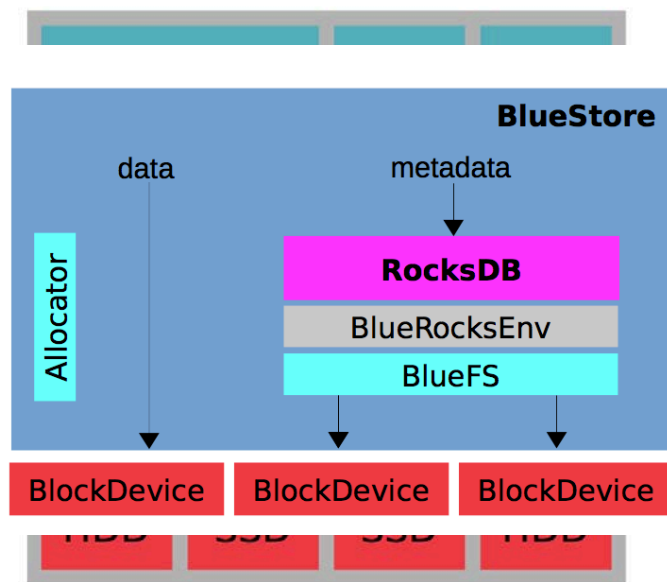


## ➤思路

- 采用写优化的系统（KV存储）将随机写变成顺序写。
- 使用基于间接索引的COW日志替代WAL日志。

## ➤社区方案——BlueStore

- NewStore=RocksDB+对象文件
- BlueStore=块设备上的NewStore
- 使用XFS/ext4来保存文件。
  - 对象数据（object data）
- 用RocksDB来维护元数据（WAL）
  - 对象元数据（onode）
  - ceph的KV omap数据
  - 数据在磁盘上的映射元数据。



# TokuDB



➤ BlueStore可以使用任何Kv存储管理元数据

➤ TokuDB对比RocksDB

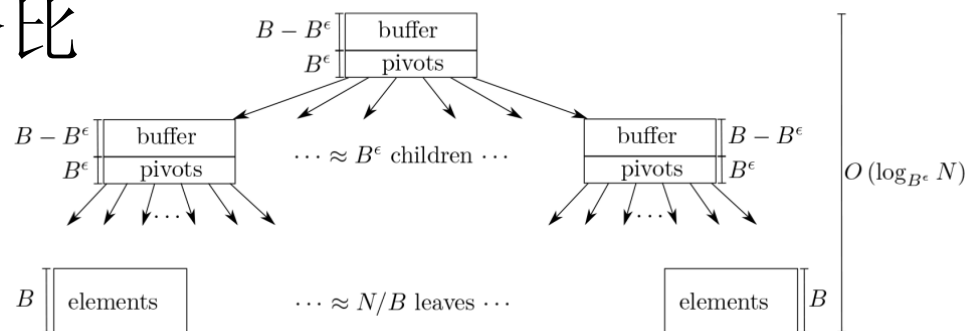
1. RocksDB——LSM Tree
2. TokuDB—— $B^\epsilon$ -Tree
3.  $B^\epsilon$ -Tree查询性能优于LSM Tree

Data Structure	Insert	Point Query		Range Query
		no Upserts	w/ Upserts	
$B^\epsilon$ -tree	$\frac{\log_B N}{\epsilon B^{1-\epsilon}}$	$\frac{\log_B N}{\epsilon}$	$\frac{\log_B N}{\epsilon}$	$\frac{\log_B N}{\epsilon} + \frac{k}{B}$
$B^\epsilon$ -tree ( $\epsilon = 1/2$ )	$\frac{\log_B N}{\sqrt{B}}$	$\log_B N$	$\log_B N$	$\log_B N + \frac{k}{B}$
B-tree	$\log_B N$	$\log_B N$	$\log_B N$	$\log_B N + \frac{k}{B}$
LSM	$\frac{\log_B N}{\epsilon B^{1-\epsilon}}$	$\frac{\log_B^2 N}{\epsilon}$	$\frac{\log_B^2 N}{\epsilon}$	$\frac{\log_B^2 N}{\epsilon} + \frac{k}{B}$
LSM+BF	$\frac{\log_B N}{\epsilon B^{1-\epsilon}}$	$\log_B N$	$\frac{\log_B^2 N}{\epsilon}$	$\frac{\log_B^2 N}{\epsilon} + \frac{k}{B}$



# $B^\epsilon$ -Tree

- 每一个节点包含多个磁盘块
- 每个节点被分为两部分
  - Pivots: 指向子节点的指针。
  - Buffer: 缓存更新和写入消息。
- Buffer是在磁盘上的
- $\epsilon$ 参数——Pivots的百分比



# B<sup>ε</sup>-Tree操作

## ➤插入/更新

1. 构建一个insert消息
2. 插入到B<sup>ε</sup>-Tree的根节点的Buffer中，写入成功返回。
3. 如果节点的Buffer满了，则将消息根据键的值写入对应的子节点的Buffer中
4. 重复第3步直到写入叶子节点。

## ➤删除

1. 在根节点插入tombstone消息
2. Buffer满后分发到子节点，直到到达叶子节点
3. 删除叶子节点对应的KV，丢弃该消息

## ➤节点的分裂与合并——与B<sup>+</sup>树类似

## ➤点查询和范围查询

- 点查询操作与B<sup>+</sup>树类似，性能也相差无几
- 范围查询由多个点查询构成，性能比B<sup>+</sup>树差
  - B<sup>+</sup>树有一个指向兄弟节点的指针



# 分形树（Fractal Tree）



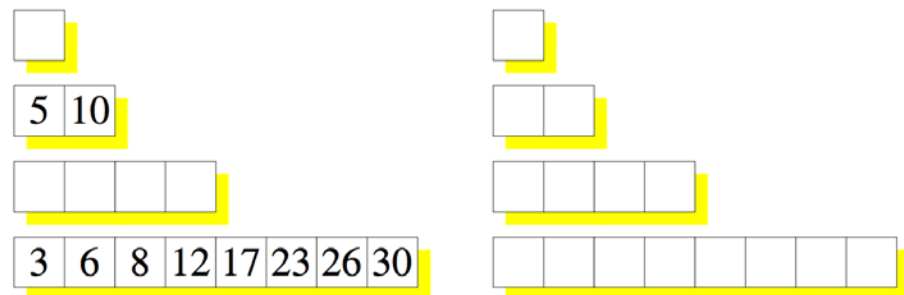
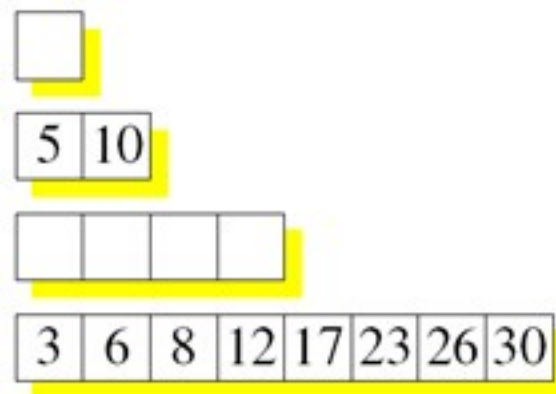
➤分形树是TokuDB中对B<sup>ε</sup>-Tree的一个开源实现

➤数据结构（简化版）

- 相邻行空间加倍
- 每一行要么全满要么全空
- 全满行的数据都是排好序的

➤辅助数组

- 辅助插入操作
- 让插入和合并由两个独立的线程操作



# 分形树的写入与合并

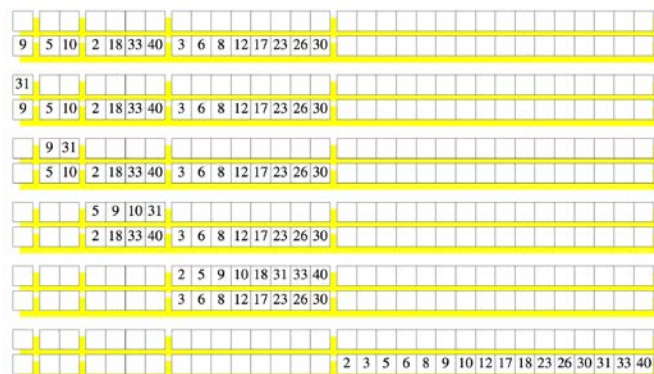
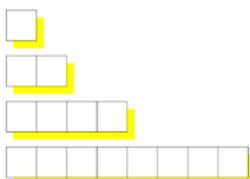
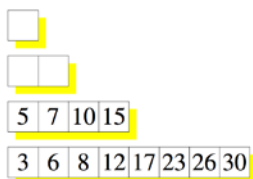
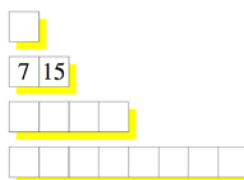
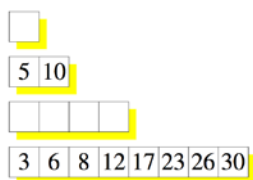
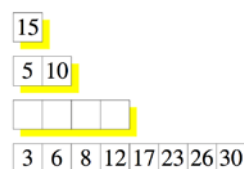
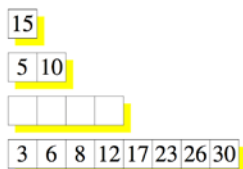
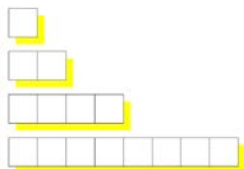
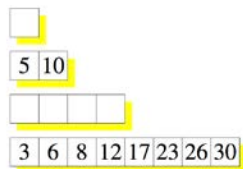
## ➤ 写入线程

- 写入第一级（根节点）
- 如果主数组空，写入主数组否则写入辅助数组
- 主数组与辅助数组必有一个为空

## ➤ 合并线程

- 当某一级的主辅数组都满的时候，将其合并写入下一级
- 写入下一级的时候如果下一级的主数组为空，写入主数组，否则写入辅数组
- 直到没有任何一级主辅数组都满

# 分形树的写入与合并



# 简单分形树的性能

## ➤ 插入

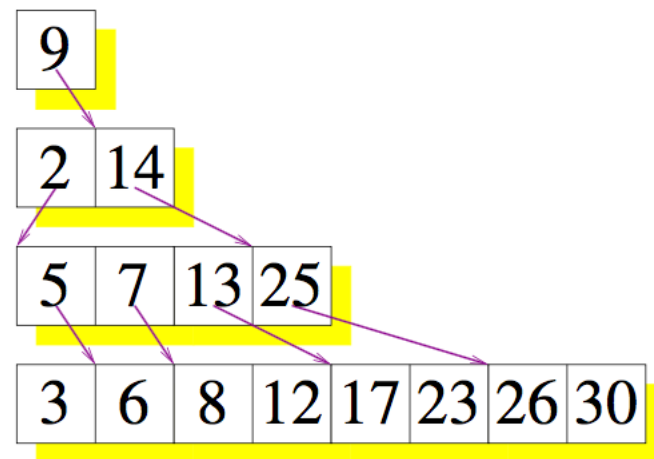
- $O(1)$

## ➤ 合并

- 合并两个大小为 $X$ 的数组 $O(X/B)$
- 每一个元素的代价是 $O(1/B)$
- 每一个元素的合并代价是 $O(\log N)$
- 平均插入代价是 $O((\log N)/B)$

## ➤ 查询性能是 $O(\log^2 N)$

- 与LSM-Tree相同
- 加入前向指针



## ➤ 前向指针

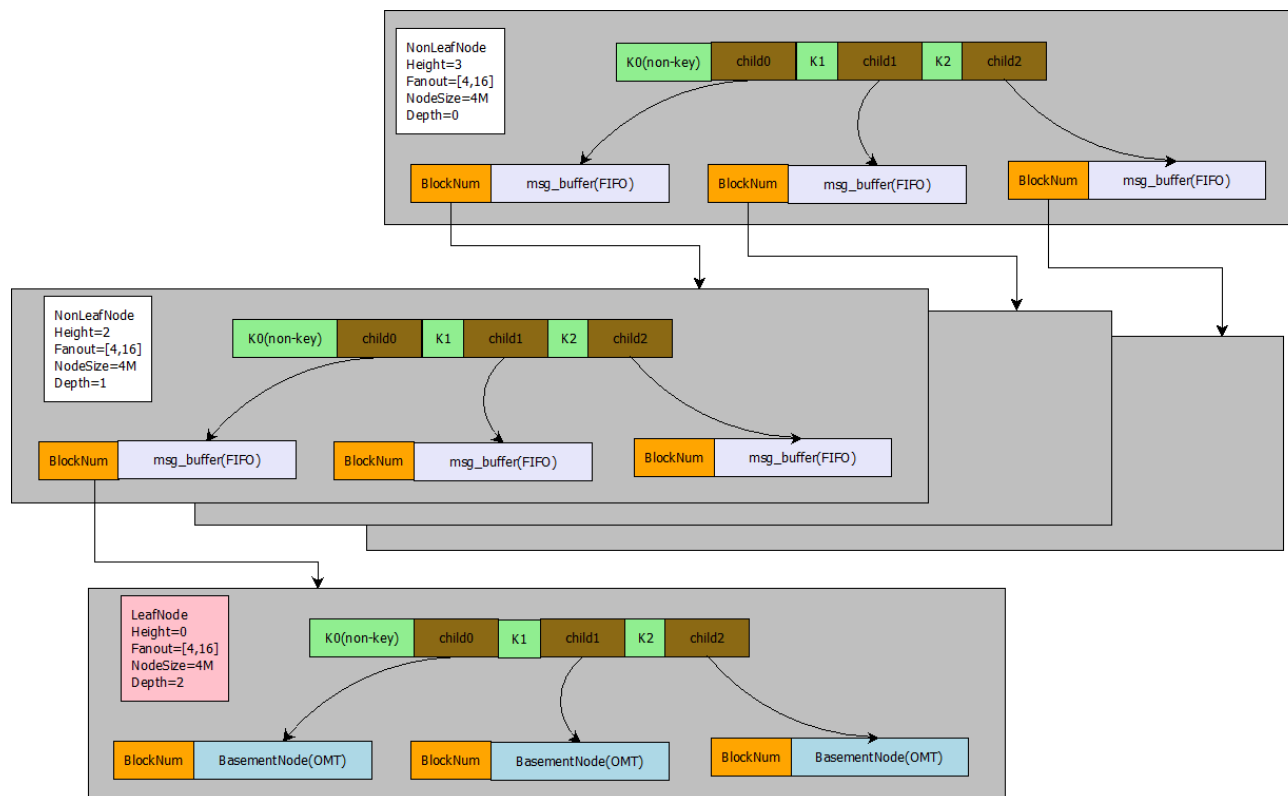
- 每一个元素都有一个指向下一级中第一个键值大于本身的元素
- 同一级采用二分查找
- 查询性能为 $O(\log N)$

# 分形树的实现



```
158 // TODO: class me up
159 struct ftnode {
160     MSN      max_msn_applied_to_node_on_disk; // max msn applied that will be written to disk
161     unsigned int 45 /* Pivot keys.
162     BLOCKNUM b 46 * Child 0's keys are <= pivotkeys[0].
163     int lay 47 * Child 1's keys are <= pivotkeys[1].
164     int lay 48 * Child 1's keys are > pivotkeys[0].
165     int lay 49 * etc
166     uint32_t b 49 */
167     int hei 50 */
168     int dir 51 class ftnode_pivot_keys {
169     uint32_t f 52 public:
170
171     // for int 53 // effect: create an empty set of pivot keys
172     // for leaf 54     233     typedef struct ftnode_child_pointer {
173     int n_child 55     234         union {
174     ftnode_piv 56     235             struct sub_block *subblock;
175     --- 57     236             struct ftnode *leaf_childinfo;
176     58     237     };
177     238     };
178     239     };
179     240     };
180     241     };
181     242     };
182     243     };
183     244     };
184     245     };
185     246     };
186     247     };
187     248     };
188     249     };
189     250     };
190     251     };
191     252     };
192     253     };
193     254     };
194     255     };
195     256     };
196     257     };
197     258     };
198     259     };
199     260     };
200     261     };
201     262     };
202     263     };
203     264     };
204     265     };
205     266     };
206     267     };
207     268     };
208     269     };
209     270     };
210     271     };
211     272     };
212     273     };
213     274     };
214     275     };
215     276     };
216     277     };
217     278     };
218     279     };
219     280     };
220     281     };
221     282     };
222     283     };
223     284     };
224     285     };
225     286     };
226     287     };
227     288     };
228     289     };
229     290     };
230     291     };
231     292     };
232     293     };
233     294     };
234     295     };
235     296     };
236     297     };
237     298     };
238     299     };
239     300     };
240     301     };
241     302     };
242     303     };
243     304     };
244     305     };
245     306     };
246     307     };
247     308     };
248     309     };
249     310     };
250     311     };
251     312     };
252     313     };
253     314     };
254     315     };
255     316     };
256     317     };
257     318     };
258     319     };
259     320     };
260     321     };
261     322     };
262     323     };
263     324     };
264     325     };
265     326     };
266     327     };
267     328     };
268     329     };
269     330     };
270     331     };
271     332     };
272     333     };
273     334     };
274     335     };
275     336     };
276     337     };
277     338     };
278     339     };
279     340     };
280     341     };
281     342     };
282     343     };
283     344     };
284     345     };
285     346     };
286     347     };
287     348     };
288     349     };
289     350     };
290     351     };
291     352     };
292     353     };
293     354     };
294     355     };
295     356     };
296     357     };
297     358     };
298     359     };
299     360     };
300     361     };
301     362     };
302     363     };
303     364     };
304     365     };
305     366     };
306     367     };
307     368     };
308     369     };
309     370     };
310     371     };
311     372     };
312     373     };
313     374     };
314     375     };
315     376     };
316     377     };
317     378     };
318     379     };
319     380     };
320     381     };
321     382     };
322     383     };
323     384     };
324     385     };
325     386     };
326     387     };
327     388     };
328     389     };
329     390     };
330     391     };
331     392     };
332     393     };
333     394     };
334     395     };
335     396     };
336     397     };
337     398     };
338     399     };
339     400     };
340     401     };
341     402     };
342     403     };
343     404     };
344     405     };
345     406     };
346     407     };
347     408     };
348     409     };
349     410     };
350     411     };
351     412     };
352     413     };
353     414     };
354     415     };
355     416     };
356     417     };
357     418     };
358     419     };
359     420     };
360     421     };
361     422     };
362     423     };
363     424     };
364     425     };
365     426     };
366     427     };
367     428     };
368     429     };
369     430     };
370     431     };
371     432     };
372     433     };
373     434     };
374     435     };
375     436     };
376     437     };
377     438     };
378     439     };
379     440     };
380     441     };
381     442     };
382     443     };
383     444     };
384     445     };
385     446     };
386     447     };
387     448     };
388     449     };
389     450     };
390     451     };
391     452     };
392     453     };
393     454     };
394     455     };
395     456     };
396     457     };
397     458     };
398     459     };
399     460     };
400     461     };
401     462     };
402     463     };
403     464     };
404     465     };
405     466     };
406     467     };
407     468     };
408     469     };
409     470     };
410     471     };
411     472     };
412     473     };
413     474     };
414     475     };
415     476     };
416     477     };
417     478     };
418     479     };
419     480     };
420     481     };
421     482     };
422     483     };
423     484     };
424     485     };
425     486     };
426     487     };
427     488     };
428     489     };
429     490     };
430     491     };
431     492     };
432     493     };
433     494     };
434     495     };
435     496     };
436     497     };
437     498     };
438     499     };
439     500     };
440     501     };
441     502     };
442     503     };
443     504     };
444     505     };
445     506     };
446     507     };
447     508     };
448     509     };
449     510     };
450     511     };
451     512     };
452     513     };
453     514     };
454     515     };
455     516     };
456     517     };
457     518     };
458     519     };
459     520     };
460     521     };
461     522     };
462     523     };
463     524     };
464     525     };
465     526     };
466     527     };
467     528     };
468     529     };
469     530     };
470     531     };
471     532     };
472     533     };
473     534     };
474     535     };
475     536     };
476     537     };
477     538     };
478     539     };
479     540     };
480     541     };
481     542     };
482     543     };
483     544     };
484     545     };
485     546     };
486     547     };
487     548     };
488     549     };
489     550     };
490     551     };
491     552     };
492     553     };
493     554     };
494     555     };
495     556     };
496     557     };
497     558     };
498     559     };
499     560     };
500     561     };
501     562     };
502     563     };
503     564     };
504     565     };
505     566     };
506     567     };
507     568     };
508     569     };
509     570     };
510     571     };
511     572     };
512     573     };
513     574     };
514     575     };
515     576     };
516     577     };
517     578     };
518     579     };
519     580     };
520     581     };
521     582     };
522     583     };
523     584     };
524     585     };
525     586     };
526     587     };
527     588     };
528     589     };
529     590     };
530     591     };
531     592     };
532     593     };
533     594     };
534     595     };
535     596     };
536     597     };
537     598     };
538     599     };
539     600     };
540     601     };
541     602     };
542     603     };
543     604     };
544     605     };
545     606     };
546     607     };
547     608     };
548     609     };
549     610     };
550     611     };
551     612     };
552     613     };
553     614     };
554     615     };
555     616     };
556     617     };
557     618     };
558     619     };
559     620     };
560     621     };
561     622     };
562     623     };
563     624     };
564     625     };
565     626     };
566     627     };
567     628     };
568     629     };
569     630     };
570     631     };
571     632     };
572     633     };
573     634     };
574     635     };
575     636     };
576     637     };
577     638     };
578     639     };
579     640     };
580     641     };
581     642     };
582     643     };
583     644     };
584     645     };
585     646     };
586     647     };
587     648     };
588     649     };
589     650     };
590     651     };
591     652     };
592     653     };
593     654     };
594     655     };
595     656     };
596     657     };
597     658     };
598     659     };
599     660     };
600     661     };
601     662     };
602     663     };
603     664     };
604     665     };
605     666     };
606     667     };
607     668     };
608     669     };
609     670     };
610     671     };
611     672     };
612     673     };
613     674     };
614     675     };
615     676     };
616     677     };
617     678     };
618     679     };
619     680     };
620     681     };
621     682     };
622     683     };
623     684     };
624     685     };
625     686     };
626     687     };
627     688     };
628     689     };
629     690     };
630     691     };
631     692     };
632     693     };
633     694     };
634     695     };
635     696     };
636     697     };
637     698     };
638     699     };
639     700     };
640     701     };
641     702     };
642     703     };
643     704     };
644     705     };
645     706     };
646     707     };
647     708     };
648     709     };
649     710     };
650     711     };
651     712     };
652     713     };
653     714     };
654     715     };
655     716     };
656     717     };
657     718     };
658     719     };
659     720     };
660     721     };
661     722     };
662     723     };
663     724     };
664     725     };
665     726     };
666     727     };
667     728     };
668     729     };
669     730     };
670     731     };
671     732     };
672     733     };
673     734     };
674     735     };
675     736     };
676     737     };
677     738     };
678     739     };
679     740     };
680     741     };
681     742     };
682     743     };
683     744     };
684     745     };
685     746     };
686     747     };
687     748     };
688     749     };
689     750     };
690     751     };
691     752     };
692     753     };
693     754     };
694     755     };
695     756     };
696     757     };
697     758     };
698     759     };
699     760     };
700     761     };
701     762     };
702     763     };
703     764     };
704     765     };
705     766     };
706     767     };
707     768     };
708     769     };
709     770     };
710     771     };
711     772     };
712     773     };
713     774     };
714     775     };
715     776     };
716     777     };
717     778     };
718     779     };
719     780     };
720     781     };
721     782     };
722     783     };
723     784     };
724     785     };
725     786     };
726     787     };
727     788     };
728     789     };
729     790     };
730     791     };
731     792     };
732     793     };
733     794     };
734     795     };
735     796     };
736     797     };
737     798     };
738     799     };
739     800     };
740     801     };
741     802     };
742     803     };
743     804     };
744     805     };
745     806     };
746     807     };
747     808     };
748     809     };
749     810     };
750     811     };
751     812     };
752     813     };
753     814     };
754     815     };
755     816     };
756     817     };
757     818     };
758     819     };
759     820     };
760     821     };
761     822     };
762     823     };
763     824     };
764     825     };
765     826     };
766     827     };
767     828     };
768     829     };
769     830     };
770     831     };
771     832     };
772     833     };
773     834     };
774     835     };
775     836     };
776     837     };
777     838     };
778     839     };
779     840     };
780     841     };
781     842     };
782     843     };
783     844     };
784     845     };
785     846     };
786     847     };
787     848     };
788     849     };
789     850     };
790     851     };
791     852     };
792     853     };
793     854     };
794     855     };
795     856     };
796     857     };
797     858     };
798     859     };
799     860     };
800     861     };
801     862     };
802     863     };
803     864     };
804     865     };
805     866     };
806     867     };
807     868     };
808     869     };
809     870     };
810     871     };
811     872     };
812     873     };
813     874     };
814     875     };
815     876     };
816     877     };
817     878     };
818     879     };
819     880     };
820     881     };
821     882     };
822     883     };
823     884     };
824     885     };
825     886     };
826     887     };
827     888     };
828     889     };
829     890     };
830     891     };
831     892     };
832     893     };
833     894     };
834     895     };
835     896     };
836     897     };
837     898     };
838     899     };
839     900     };
840     901     };
841     902     };
842     903     };
843     904     };
844     905     };
845     906     };
846     907     };
847     908     };
848     909     };
849     910     };
850     911     };
851     912     };
852     913     };
853     914     };
854     915     };
855     916     };
856     917     };
857     918     };
858     919     };
859     920     };
860     921     };
861     922     };
862     923     };
863     924     };
864     925     };
865     926     };
866     927     };
867     928     };
868     929     };
869     930     };
870     931     };
871     932     };
872     933     };
873     934     };
874     935     };
875     936     };
876     937     };
877     938     };
878     939     };
879     940     };
880     941     };
881     942     };
882     943     };
883     944     };
884     945     };
885     946     };
886     947     };
887     948     };
888     949     };
889     950     };
890     951     };
891     952     };
892     953     };
893     954     };
894     955     };
895     956     };
896     957     };
897     958     };
898     959     };
899     960     };
900     961     };
901     962     };
902     963     };
903     964     };
904     965     };
905     966     };
906     967     };
907     968     };
908     969     };
909     970     };
910     971     };
911     972     };
912     973     };
913     974     };
914     975     };
915     976     };
916     977     };
917     978     };
918     979     };
919     980     };
920     981     };
921     982     };
922     983     };
923     984     };
924     985     };
925     986     };
926     987     };
927     988     };
928     989     };
929     990     };
930     991     };
931     992     };
932     993     };
933     994     };
934     995     };
935     996     };
936     997     };
937     998     };
938     999     };
939     1000     };
940     1001     };
941     1002     };
942     1003     };
943     1004     };
944     1005     };
945     1006     };
946     1007     };
947     1008     };
948     1009     };
949     1010     };
950     1011     };
951     1012     };
952     1013     };
953     1014     };
954     1015     };
955     1016     };
956     1017     };
957     1018     };
958     1019     };
959     1020     };
960     1021     };
961     1022     };
962     1023     };
963     1024     };
964     1025     };
965     1026     };
966     1027     };
967     1028     };
968     1029     };
969     1030     };
970     1031     };
971     1032     };
972     1033     };
973     1034     };
974     1035     };
975     1036     };
976     1037     };
977     1038     };
978     1039     };
979     1040     };
980     1041     };
981     1042     };
982     1043     };
983     1044     };
984     1045     };
985     1046     };
986     1047     };
987     1048     };
988     1049     };
989     1050     };
990     1051     };
991     1052     };
992     1053     };
993     1054     };
994     1055     };
995     1056     };
996     1057     };
997     1058     };
998     1059     };
999     1060     };
1000     1061     };
1001     1062     };
1002     1063     };
1003     1064     };
1004     1065     };
1005     1066     };
1006     1067     };
1007     1068     };
1008     1069     };
1009     1070     };
1010     1071     };
1011     1072     };
1012     1073     };
1013     1074     };
1014     1075     };
1015     1076     };
1016     1077     };
1017     1078     };
1018     1079     };
1019     1080     };
1020     1081     };
1021     1082     };
1022     1083     };
1023     1084     };
1024     1085     };
1025     1086     };
1026     1087     };
1027     1088     };
1028     1089     };
1029     1090     };
1030     1091     };
1031     1092     };
1032     1093     };
1033     1094     };
1034     1095     };
1035     1096     };
1036     1097     };
1037     1098     };
1038     1099     };
1039     1100     };
1040     1101     };
1041     1102     };
1042     1103     };
1043     1104     };
1044     1105     };
1045     1106     };
1046     1107     };
1047     1108     };
1048     1109     };
1049     1110     };
1050     1111     };
1051     1112     };
1052     1113     };
1053     1114     };
1054     1115     };
1055     1116     };
1056     1117     };
1057     1118     };
1058     1119     };
1059     1120     };
1060     1121     };
1061     1122     };
1062     1123     };
1063     1124     };
1064     1125     };
1065     1126     };
1066     1127     };
1067     1128     };
1068     1129     };
1069     1130     };
1070     1131     };
1071     1132     };
1072     1133     };
1073     1134     };
1074     1135     };
1075     1136     };
1076     1137     };
1077     1138     };
1078     1139     };
1079     1140     };
1080     1141     };
1081     1142     };
1082     1143     };
1083     1144     };
1084     1145     };
1085     1146     };
1086     1147     };
1087     1148     };
1088     1149     };
1089     1150     };
1090     1151     };
1091     1152     };
1092     1153     };
1093     1154     };
1094     1155     };
1095     1156     };
1096     1157     };
1097     1158     };
1098     1159     };
1099     1160     };
1100     1161     };
1101     1162     };
1102     1163     };
1103     1164     };
1104     1165     };
1105     1166     };
1106     1167     };
1107     1168     };
1108     1169     };
1109     1170     };
1110     1171     };
1111     1172     };
1112     1173     };
1113     1174     };
1114     1175     };
1115     1176     };
1116     1177     };
1117     1178     };
1118     1179     };
1119     1180     };
1120     1181     };
1121     1182     };
1122     1183     };
1123     1184     };
1124     1185     };
1125     1186     };
1126     1187     };
1127     1188     };
1128     1189     };
1129     1190     };
1130     1191     };
1131     1192     };
1132     1193     };
1133     1194     };
1134     1195     };
1135     1196     };
1136     1197     };
1137     1198     };
1138     1199     };
1139     1200     };
1140     1201     };
1141     1202     };
1142     1203     };
1143     1204     };
1144     1205     };
1145     1206     };
1146     1207     };
1147     1208     };
1148     1209     };
1149     1210     };
1150     1211     };
1151     1212     };
1152     1213     };
1153     1214     };
1154     1215     };
1155     1216     };
1156     1217     };
1157     1218     };
1158     1219     };
1159     1220     };
1160     1221     };
1161     1222     };
1162     1223     };
1163     1224     };
1164     1225     };
1165     1226     };
1166     1227     };
1167     1228     };
1168     1229     };
1169     1230     };
1170     1231     };
1171     1232     };
1172     1233     };
1173     1234     };
1174     1235     };
1175     1236     };
1176     1237     };
1177     1238     };
1178     1239     };
1179     1240     };
1180     1241     };
1181     1242     };
1182     1243     };
1183     1244     };
1184     1245     };
1185     1246     };
1186     1247     };
1187     1248     };
1188     1249     };
1189     1250     };
1190     1251     };
1191     1252     };
1192     1253     };
1193     1254     };
1194     1255     };
1195     1256     };
1196     1257     };
1197     1258     };
1198     1259     };
1199     1260     };
1200     1261     };
1201     1262     };
1202     1263     };
1203     1264     };
1204     1265     };
1205     1266     };
1206     1267     };
1207     1268     };
1208     1269     };
1209     1270     };
1210     1271     };
1211     1272     };
1212     1273     };
1213     1274     };
1214    
```

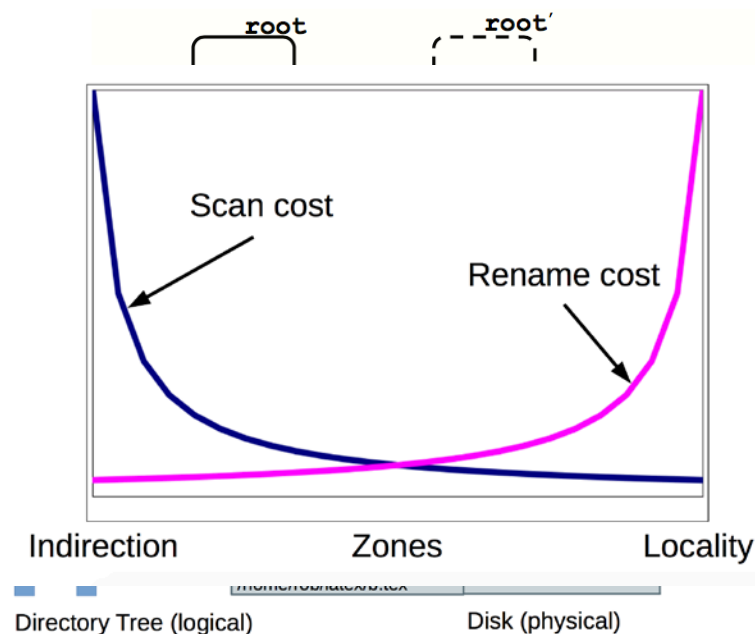
# 分形树的实现



# BetrFS



- 使用分形树
- late-binding 日志（BlueStore 中的延迟写日志？）
- 分区（zoning）
- Range Deletion
  - 删除消息的广播



# 未来工作

---



- 用TokuDB替代RocksDB
- 测试性能，TokuDB和RocksDB的性能对比
- 引入late-binding日志和分区



谢谢！

Q&A