3fs实验

实验结果

BigDataBench-MPI-Sort对比实验

和其他分布式文件系统对比：

benchmark	3FS运行时间	HDFS运行时间	Lustre系统运行时间
MPI-Sort	33s	40s	5s

使用fio工具进行实验

使用fio工具的原因是3fs本身不使用cache，可以通过fio工具进行直接测试（-direct=1参数）。
读取结果：

```
顺序读：
READ: bw=43.8MiB/s (45.9MB/s), 43.8MiB/s-43.8MiB/s (45.9MB/s-45.9MB/s), io=7880MiB (8263MB), run=180073-180073msec
随机读：
READ: bw=40.4MiB/s (42.4MB/s), 40.4MiB/s-40.4MiB/s (42.4MB/s-42.4MB/s), io=7280MiB (7634MB), run=180135-180135msec
```

BigDataBench-MPI-Sort实验

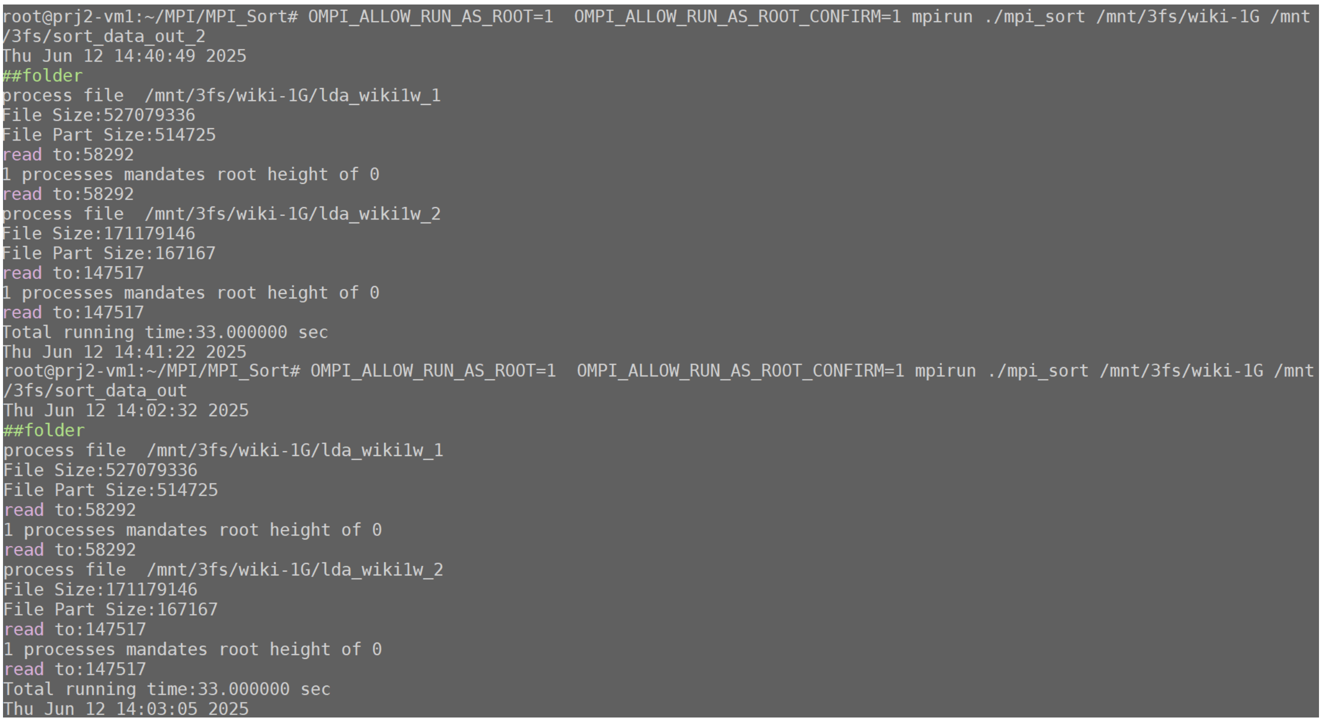
排序时间统计

单机上排序时间：33sec

【五次实验取众数,33sec-36sec之间】

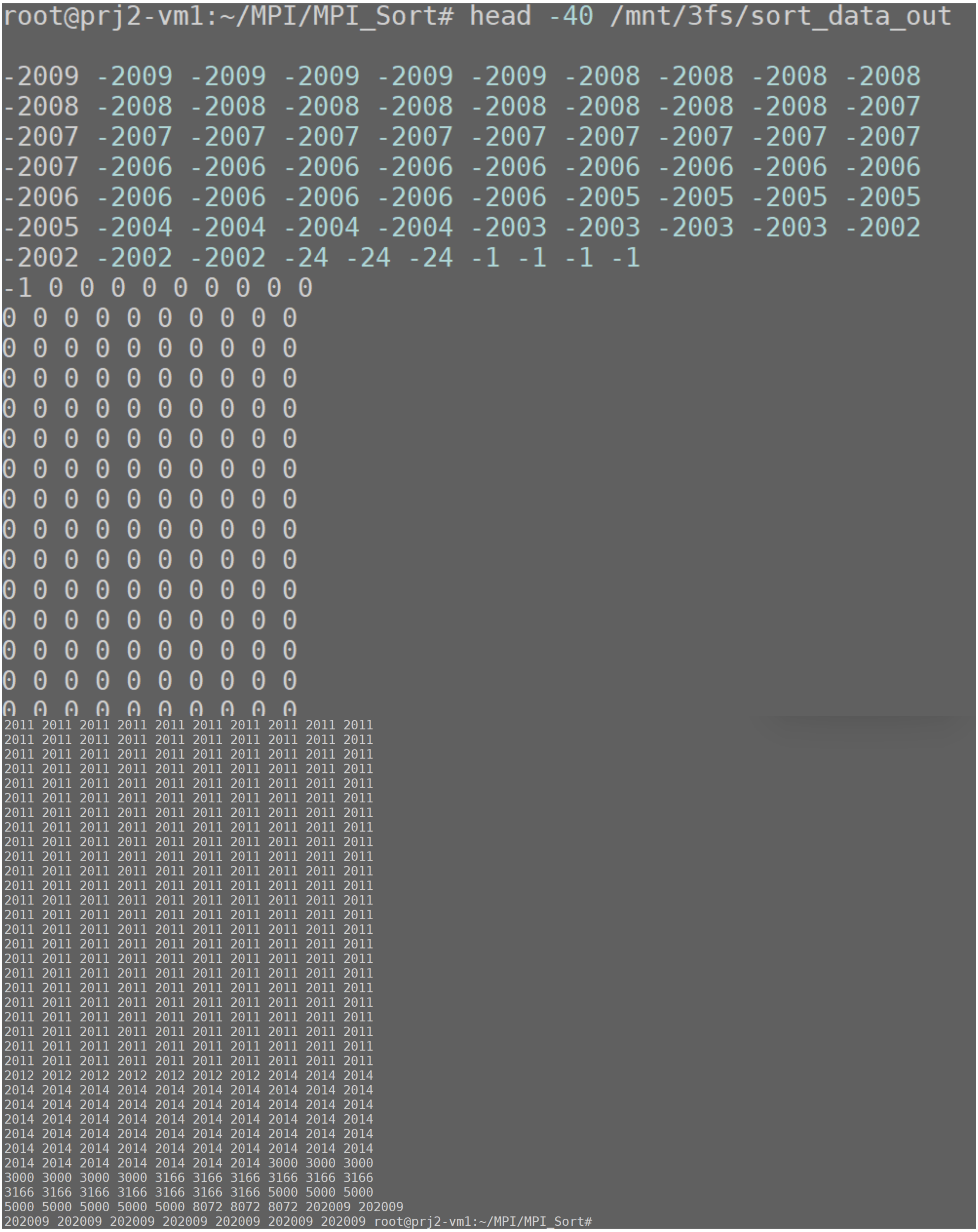
```
 OMPI_ALLOW_RUN_AS_ROOT=1  OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun ./mpi_sort /mnt/3fs/wiki-1G /mnt/3fs/sort_data_out
```

实验结果：



```
root@prj2-vm1:~/MPI/MPI_Sort# OMPI_ALLOW_RUN_AS_ROOT=1  OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun ./mpi_sort /mnt/3fs/wiki-1G /mnt
/3fs/sort_data_out_2
Thu Jun 12 14:40:49 2025
##folder
process file  /mnt/3fs/wiki-1G/lda_wiki1w_1
File Size:527079336
File Part Size:514725
read to:58292
1 processes mandates root height of 0
read to:58292
process file  /mnt/3fs/wiki-1G/lda_wiki1w_2
File Size:171179146
File Part Size:167167
read to:147517
1 processes mandates root height of 0
read to:147517
Total running time:33.000000 sec
Thu Jun 12 14:41:22 2025
root@prj2-vm1:~/MPI/MPI_Sort# OMPI_ALLOW_RUN_AS_ROOT=1  OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun ./mpi_sort /mnt/3fs/wiki-1G /mnt
/3fs/sort_data_out
Thu Jun 12 14:02:32 2025
##folder
process file  /mnt/3fs/wiki-1G/lda_wiki1w_1
File Size:527079336
File Part Size:514725
read to:58292
1 processes mandates root height of 0
read to:58292
process file  /mnt/3fs/wiki-1G/lda_wiki1w_2
File Size:171179146
File Part Size:167167
read to:147517
1 processes mandates root height of 0
read to:147517
Total running time:33.000000 sec
Thu Jun 12 14:03:05 2025
```

排序结果：



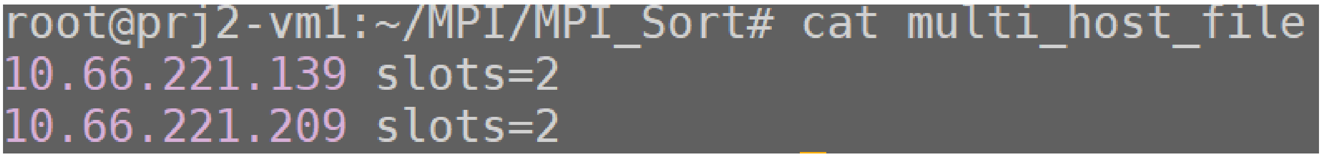
```
root@prj2-vm1:~/MPI/MPI_Sort# head -40 /mnt/3fs/sort_data_out

-2009 -2009 -2009 -2009 -2009 -2009 -2008 -2008 -2008 -2008
-2008 -2008 -2008 -2008 -2008 -2008 -2008 -2008 -2008 -2007
-2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007
-2007 -2006 -2006 -2006 -2006 -2006 -2006 -2006 -2006 -2006
-2006 -2006 -2006 -2006 -2006 -2006 -2005 -2005 -2005 -2005
-2005 -2004 -2004 -2004 -2004 -2003 -2003 -2003 -2003 -2002
-2002 -2002 -2002 -24 -24 -24 -1 -1 -1 -1
-1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2012 2012 2012 2012 2012 2012 2012 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 3000 3000 3000
3000 3000 3000 3000 3166 3166 3166 3166 3166 3166
3166 3166 3166 3166 3166 3166 3166 5000 5000 5000
5000 5000 5000 5000 5000 8072 8072 8072 202009 202009
202009 202009 202009 202009 202009 202009 202009 root@prj2-vm1:~/MPI/MPI_Sort#
```

双机上多进程协同排序结果：122sec

【配置略等同lustre，但是对3fs来说没必要】

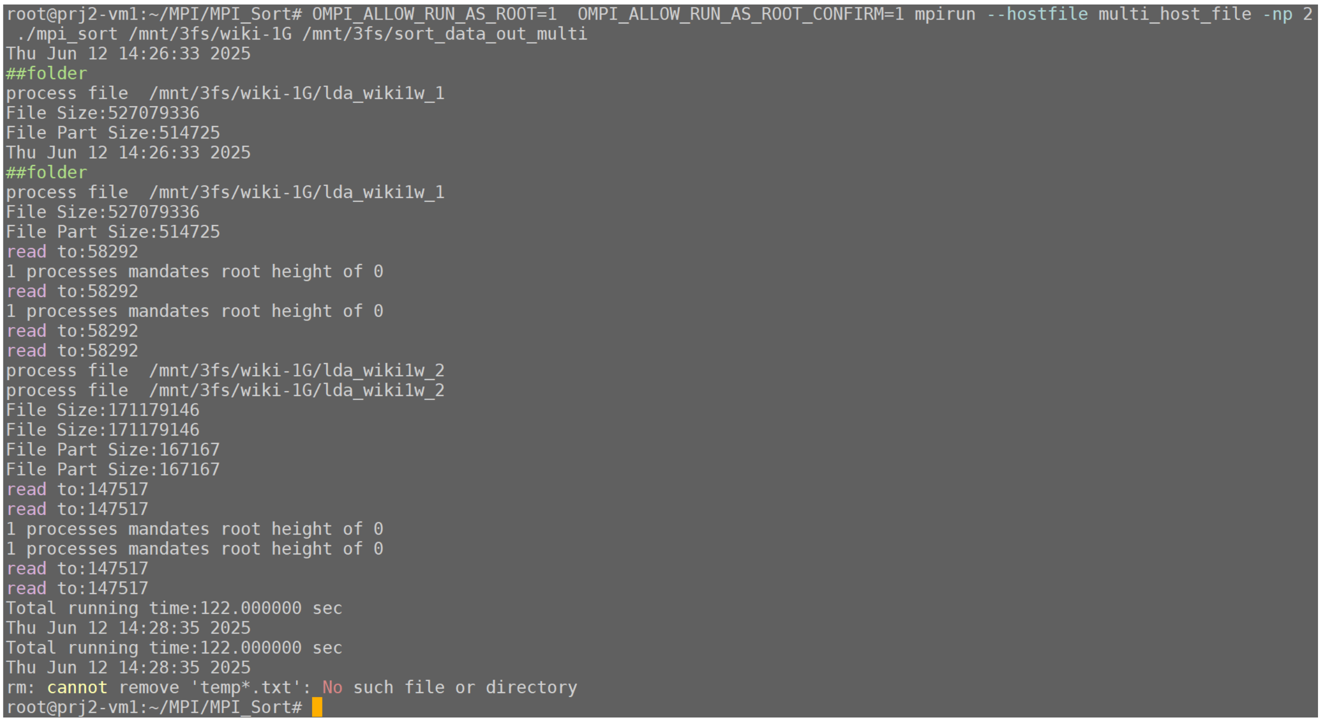
```
OMPI_ALLOW_RUN_AS_ROOT=1  OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun --
hostfile multi_host_file -np 2 ./mpi_sort /mnt/3fs/wiki-1G
/mnt/3fs/sort_data_out_multi
```

multi_host_file内容：



```
root@prj2-vm1:~/MPI/MPI_Sort# cat multi_host_file
10.66.221.139 slots=2
10.66.221.209 slots=2
```

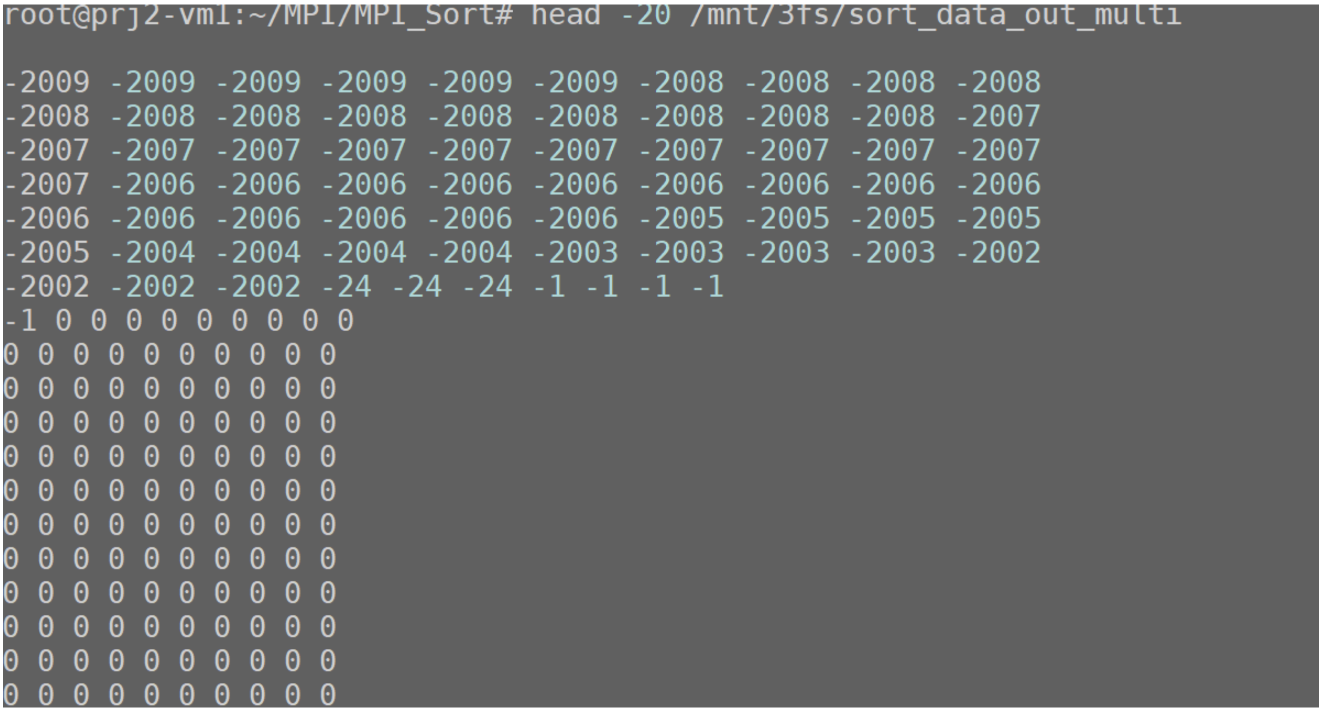
排序时间：



```
root@prj2-vm1:~/MPI/MPI_Sort# OMPI_ALLOW_RUN_AS_ROOT=1  OMPI_ALLOW_RUN_AS_ROOT_CONFIRM=1 mpirun --hostfile multi_host_file -np 2
 ./mpi_sort /mnt/3fs/wiki-1G /mnt/3fs/sort_data_out_multi
Thu Jun 12 14:26:33 2025
##folder
process file  /mnt/3fs/wiki-1G/lda_wiki1w_1
File Size:527079336
File Part Size:514725
Thu Jun 12 14:26:33 2025
##folder
process file  /mnt/3fs/wiki-1G/lda_wiki1w_1
File Size:527079336
File Part Size:514725
read to:58292
1 processes mandates root height of 0
read to:58292
1 processes mandates root height of 0
read to:58292
read to:58292
process file  /mnt/3fs/wiki-1G/lda_wiki1w_2
process file  /mnt/3fs/wiki-1G/lda_wiki1w_2
File Size:171179146
File Size:171179146
File Part Size:167167
File Part Size:167167
read to:147517
read to:147517
1 processes mandates root height of 0
1 processes mandates root height of 0
read to:147517
read to:147517
Total running time:122.000000 sec
Thu Jun 12 14:28:35 2025
Total running time:122.000000 sec
Thu Jun 12 14:28:35 2025
rm: cannot remove 'temp*.txt': No such file or directory
root@prj2-vm1:~/MPI/MPI_Sort# 
```

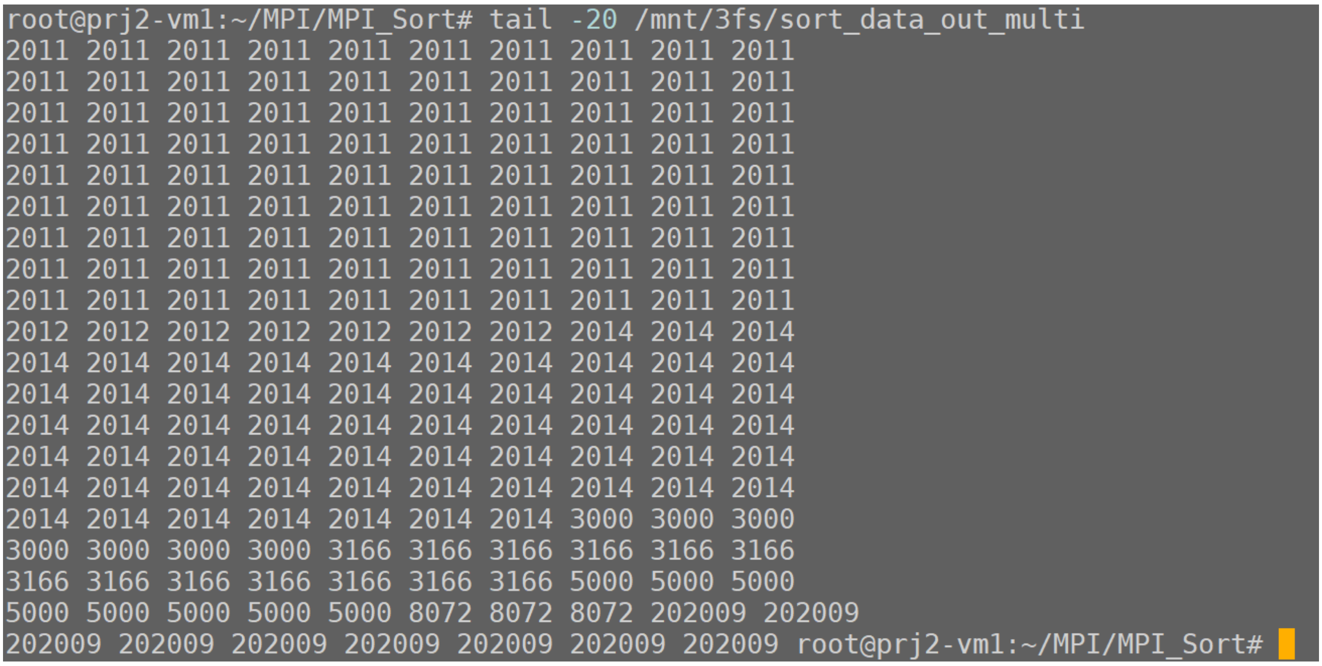
排序结果：

```
head -20 /mnt/3fs/sort_data_out_multi
```



```
root@prj2-vm1:~/MPI/MPI_Sort# head -20 /mnt/3fs/sort_data_out_multi

-2009 -2009 -2009 -2009 -2009 -2009 -2008 -2008 -2008 -2008
-2008 -2008 -2008 -2008 -2008 -2008 -2008 -2008 -2008 -2007
-2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007 -2007
-2007 -2006 -2006 -2006 -2006 -2006 -2006 -2006 -2006 -2006
-2006 -2006 -2006 -2006 -2006 -2006 -2005 -2005 -2005 -2005
-2005 -2004 -2004 -2004 -2004 -2003 -2003 -2003 -2003 -2002
-2002 -2002 -2002 -24 -24 -24 -1 -1 -1 -1
-1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
```

```
tail -20 /mnt/3fs/sort_data_out_multi
```



```
root@prj2-vm1:~/MPI/MPI_Sort# tail -20 /mnt/3fs/sort_data_out_multi
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2011 2011 2011 2011 2011 2011 2011 2011 2011 2011
2012 2012 2012 2012 2012 2012 2012 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 2014 2014 2014
2014 2014 2014 2014 2014 2014 2014 3000 3000 3000
3000 3000 3000 3000 3166 3166 3166 3166 3166 3166
3166 3166 3166 3166 3166 3166 3166 5000 5000 5000
5000 5000 5000 5000 5000 8072 8072 8072 202009 202009
202009 202009 202009 202009 202009 202009 202009 root@prj2-vm1:~/MPI/MPI_Sort# 
```

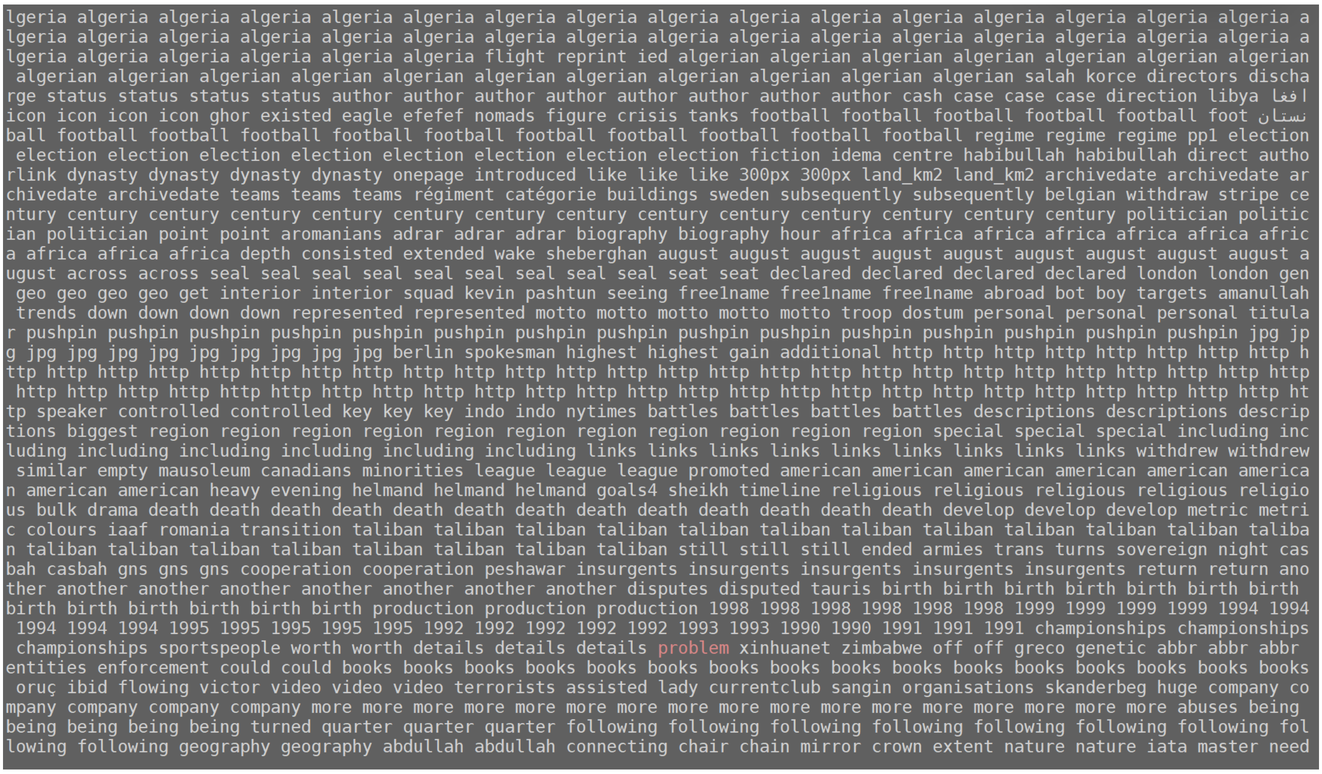
猜想：

受网络影响较大。

关于BigBench-MPI-Sort的配置

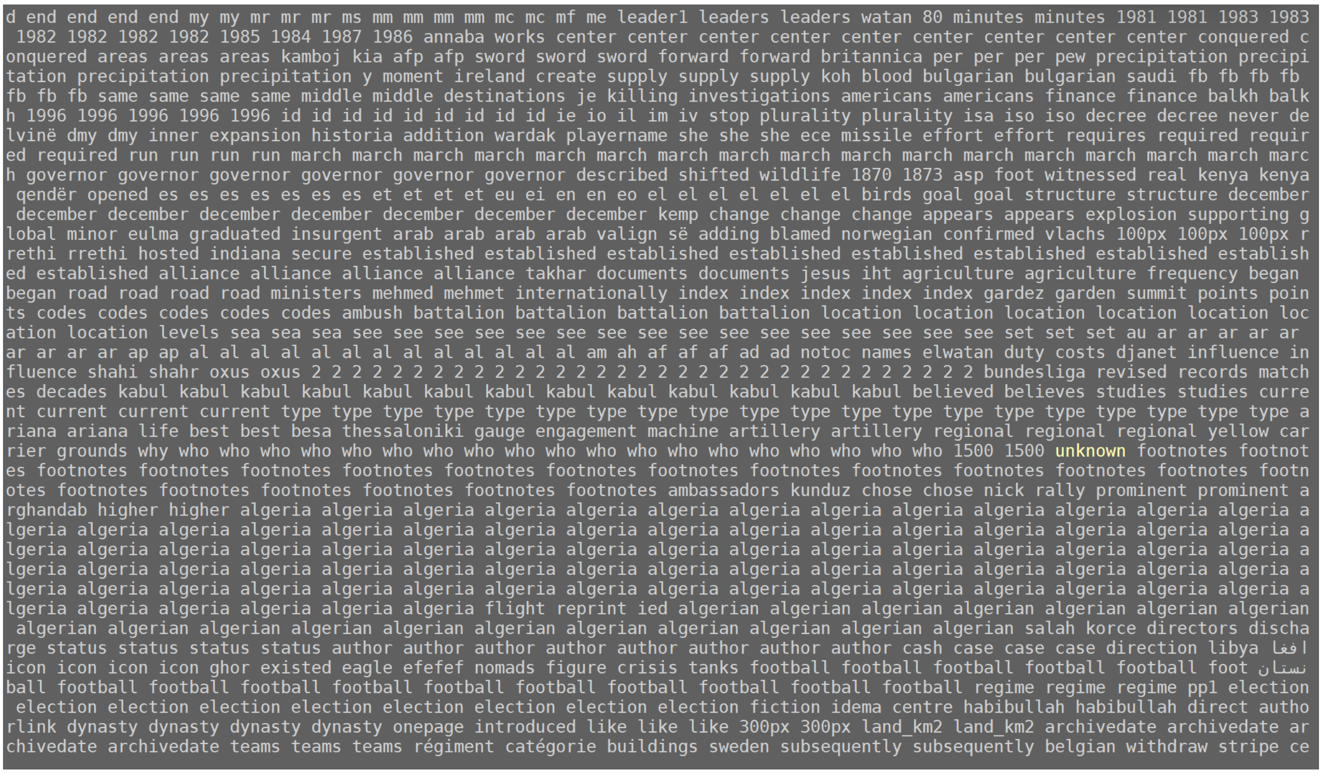
BigDataBench的MPI-Sort生成的是文本+数据

```
echo "Preparing MicroBenchmarks data dir"

WORK_DIR="/mnt/3fs"

echo "WORK_DIR=$WORK_DIR data will be put in $WORK_DIR"

cd ../../BigDataGeneratorSuite/Text_datagen/

echo "print data size GB :"
read GB
a=${GB}
L=$((a * 2))
./gen_text_data.sh lda_wiki1w $L 8000 10000 ${WORK_DIR}/wiki-$a"G"
```



```
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria flight reprint ied algerian algerian algerian algerian algerian algerian algerian
 algerian algerian algerian algerian algerian algerian algerian algerian algerian algerian algerian salah korce directors discha
rge status status status status author author author author author author author author cash case case case direction libya افغا
icon icon icon icon ghor existed eagle efefef nomads figure crisis tanks football football football football football foot نستان
ball football football football football football football football football football football regime regime regime pp1 election
 election election election election election election election election fiction idema centre habibullah habibullah direct autho
rlink dynasty dynasty dynasty dynasty onepage introduced like like like 300px 300px land_km2 land_km2 archivedate archivedate ar
chivedate archivedate teams teams teams régiment catégorie buildings sweden subsequently subsequently belgian withdraw stripe ce
ntury century century century century century century century century century century century century century politician politic
ian politician point point aromanians adrar adrar adrar biography biography hour africa africa africa africa africa africa afric
a africa africa africa depth consisted extended wake sheberghan august august august august august august august august august a
ugust across across seal seal seal seal seal seal seal seal seal seat seat declared declared declared declared london london gen
 geo geo geo geo get interior interior squad kevin pashtun seeing free1name free1name free1name abroad bot boy targets amanullah
 trends down down down down represented represented motto motto motto motto motto troop dostum personal personal personal titula
r pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin pushpin jpg jp
g jpg jpg jpg jpg jpg jpg jpg jpg jpg berlin spokesman highest highest gain additional http http http http http http http http h
ttp http http http http http http http http http http http http http http http http http http http http http http http http http
 http http http http http http http http http http http http http http http http http http http http http http http http http ht
tp speaker controlled controlled key key key indo indo nytimes battles battles battles battles descriptions descriptions descrip
tions biggest region region region region region region region region region region region special special special including inc
luding including including including including including links links links links links links links links links withdrew withdrew
 similar empty mausoleum canadians minorities league league league promoted american american american american american america
n american american heavy evening helmand helmand helmand goals4 sheikh timeline religious religious religious religious religio
us bulk drama death death death death death death death death death death death death death develop develop develop metric metri
c colours iaaf romania transition taliban taliban taliban taliban taliban taliban taliban taliban taliban taliban taliban taliba
n taliban taliban taliban taliban taliban taliban taliban taliban still still still ended armies trans turns sovereign night cas
bah casbah gns gns gns cooperation cooperation peshawar insurgents insurgents insurgents insurgents insurgents return return ano
ther another another another another another another another disputes disputed tauris birth birth birth birth birth birth birth
birth birth birth birth birth birth production production production 1998 1998 1998 1998 1998 1998 1999 1999 1999 1999 1994 1994
 1994 1994 1994 1995 1995 1995 1995 1995 1992 1992 1992 1992 1992 1993 1993 1990 1990 1991 1991 1991 championships championships
 championships sportspeople worth worth details details details problem xinhuanet zimbabwe off off greco genetic abbr abbr abbr
entities enforcement could could books books books books books books books books books books books books books books books books
 oruç ibid flowing victor video video video terrorists assisted lady currentclub sangin organisations skanderbeg huge company co
mpany company company company more more more more more more more more more more more more more more more more more abuses being
being being being being turned quarter quarter quarter following following following following following following following fol
lowing following geography geography abdullah abdullah connecting chair chain mirror crown extent nature nature iata master need
```

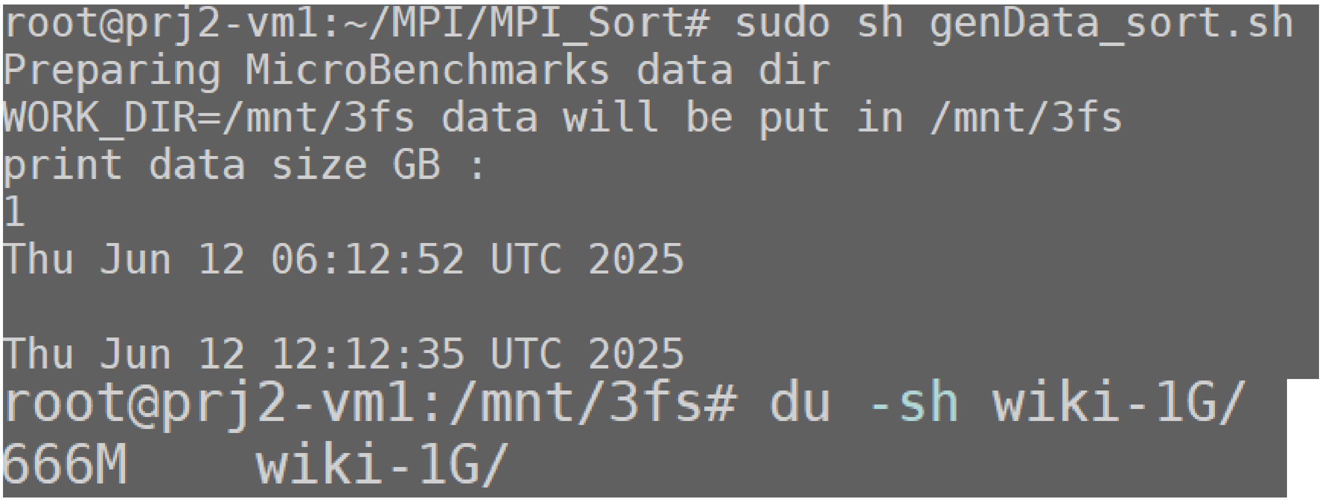
BigDataBench数据生成结果：



```
d end end end end my my mr mr mr ms mm mm mm mm mc mc mf me leader1 leaders leaders watan 80 minutes minutes 1981 1981 1983 1983
 1982 1982 1982 1982 1985 1984 1987 1986 annaba works center center center center center center center center center conquered c
onquered areas areas areas kamboj kia afp afp sword sword sword forward forward britannica per per per pew precipitation precipi
tation precipitation precipitation y moment ireland create supply supply supply koh blood bulgarian bulgarian saudi fb fb fb fb
fb fb fb same same same same middle middle destinations je killing investigations americans americans finance finance balkh balk
h 1996 1996 1996 1996 1996 id id id id id id id id id ie io il im iv stop plurality plurality isa iso iso decree decree never de
lvinë dmy dmy inner expansion historia addition wardak playername she she she ece missile effort effort requires required requir
ed required run run run run march march march march march march march march march march march march march march march march marc
h governor governor governor governor governor governor described shifted wildlife 1870 1873 asp foot witnessed real kenya kenya
 qendër opened es es es es es es es et et et et eu ei en en eo el el el el el el el birds goal goal structure structure december
 december december december december december december december kemp change change change appears appears explosion supporting g
lobal minor eulma graduated insurgent arab arab arab arab valign së adding blamed norwegian confirmed vlachs 100px 100px 100px r
rethi rrethi hosted indiana secure established established established established established established established establish
ed established alliance alliance alliance alliance takhar documents documents jesus iht agriculture agriculture frequency began
began road road road road ministers mehmed mehmet internationally index index index index index gardez garden summit points poin
ts codes codes codes codes codes ambush battalion battalion battalion battalion location location location location location loc
ation location levels sea sea sea see see see see see see see see see see see see see see see see set set set au ar ar ar ar ar
ar ar ar ar ap ap al al al al al al al al al al al al al am ah af af af ad ad notoc names elwatan duty costs djanet influence in
fluence shahi shahr oxus oxus 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 bundesliga revised records match
es decades kabul kabul kabul kabul kabul kabul kabul kabul kabul kabul kabul kabul kabul believed believes studies studies curre
nt current current current type type type type type type type type type type type type type type type type type type type type a
riana ariana life best best besa thessaloniki gauge engagement machine artillery artillery regional regional regional yellow car
rier grounds why who who who who who who who who who who who who who who who who who who who 1500 1500 unknown footnotes footnot
es footnotes footnotes footnotes footnotes footnotes footnotes footnotes footnotes footnotes footnotes footnotes footnotes footn
otes footnotes footnotes footnotes footnotes footnotes footnotes ambassadors kunduz chose chose nick rally prominent prominent a
rghandab higher higher algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria algeria a
lgeria algeria algeria algeria algeria algeria flight reprint ied algerian algerian algerian algerian algerian algerian algerian
 algerian algerian algerian algerian algerian algerian algerian algerian algerian algerian algerian salah korce directors discha
rge status status status status author author author author author author author author cash case case case direction libya افغا
icon icon icon icon ghor existed eagle efefef nomads figure crisis tanks football football football football football foot نستان
ball football football football football football football football football football football regime regime regime pp1 election
 election election election election election election election election fiction idema centre habibullah habibullah direct autho
rlink dynasty dynasty dynasty dynasty onepage introduced like like like 300px 300px land_km2 land_km2 archivedate archivedate ar
chivedate archivedate teams teams teams régiment catégorie buildings sweden subsequently subsequently belgian withdraw stripe ce
```

其他问题

排序文本预期生成的是1个G，但是为什么输出的是666M?

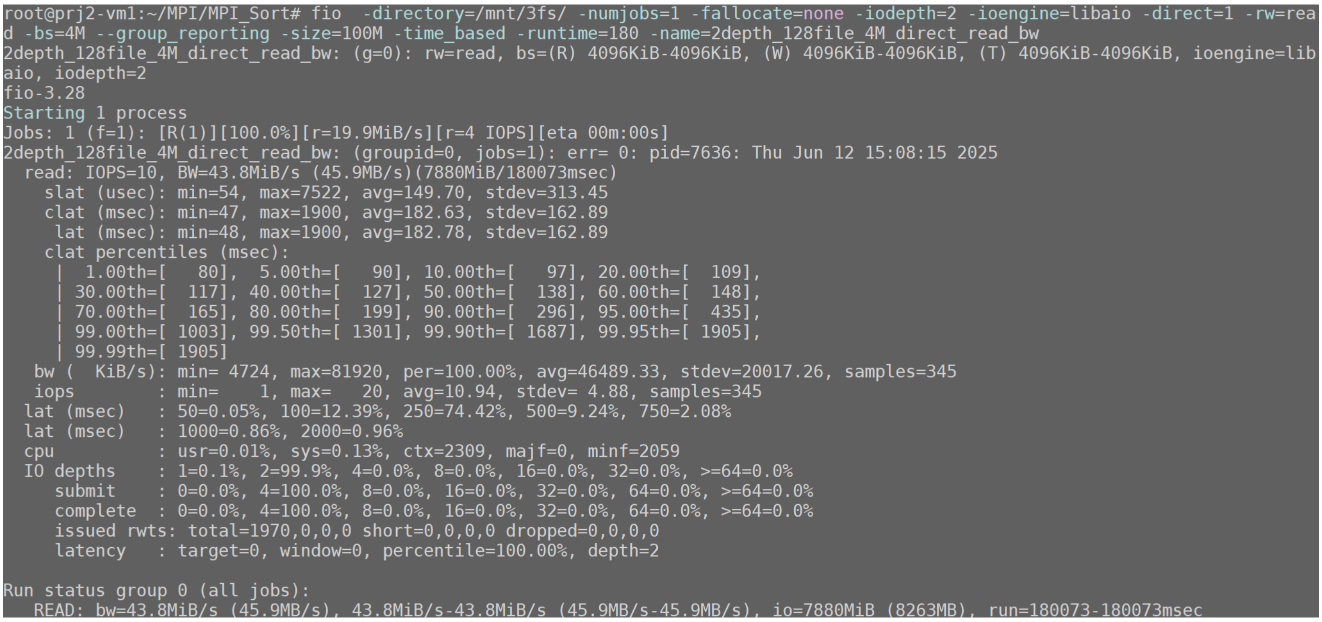


```
root@prj2-vm1:~/MPI/MPI_Sort# sudo sh genData_sort.sh
Preparing MicroBenchmarks data dir
WORK_DIR=/mnt/3fs data will be put in /mnt/3fs
print data size GB :
1
Thu Jun 12 06:12:52 UTC 2025

Thu Jun 12 12:12:35 UTC 2025
root@prj2-vm1:/mnt/3fs# du -sh wiki-1G/
666M    wiki-1G/
```

fio实验

顺序读:

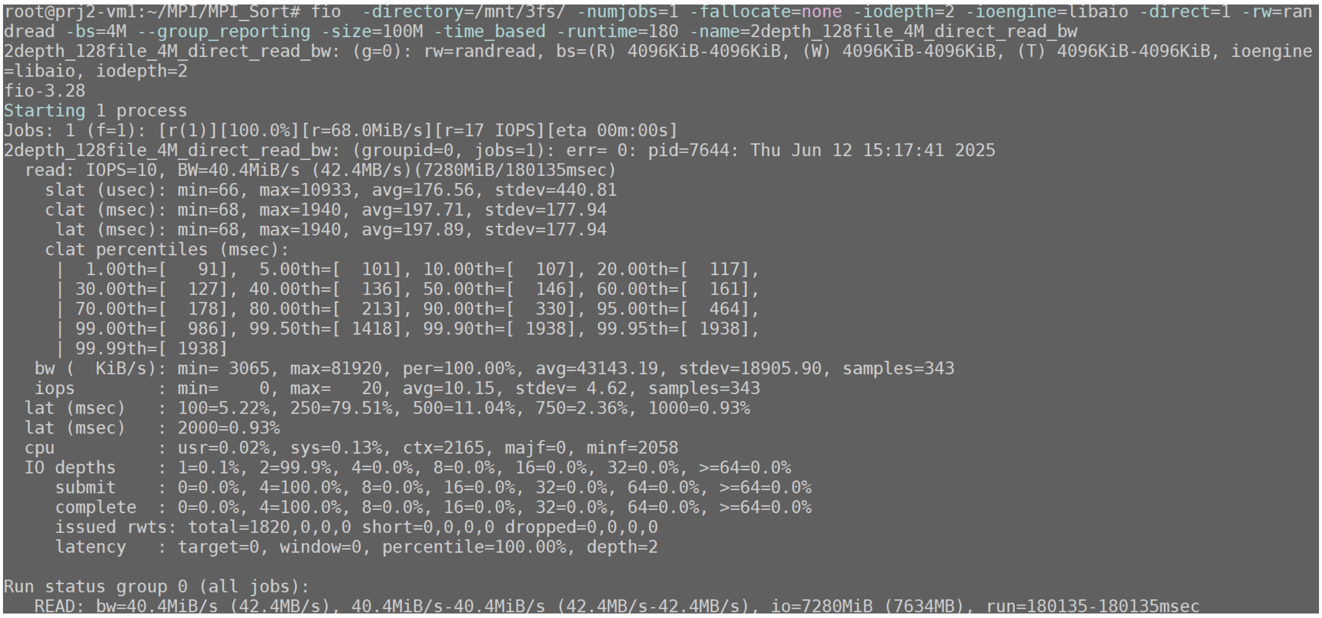
```
fio  -directory=/mnt/3fs/ -numjobs=1 -fallocate=none -iodepth=2 -
ioengine=libaio -direct=1 -rw=read -bs=4M --group_reporting -size=100M -
time_based -runtime=180 -name=2depth_128file_4M_direct_read_bw
```



```
root@prj2-vm1:~/MPI/MPI_Sort# fio  -directory=/mnt/3fs/ -numjobs=1 -fallocate=none -iodepth=2 -ioengine=libaio -direct=1 -rw=rea
d -bs=4M --group_reporting -size=100M -time_based -runtime=180 -name=2depth_128file_4M_direct_read_bw
2depth_128file_4M_direct_read_bw: (g=0): rw=read, bs=(R) 4096KiB-4096KiB, (W) 4096KiB-4096KiB, (T) 4096KiB-4096KiB, ioengine=lib
aio, iodepth=2
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [R(1)][100.0%][r=19.9MiB/s][r=4 IOPS][eta 00m:00s]
2depth_128file_4M_direct_read_bw: (groupid=0, jobs=1): err= 0: pid=7636: Thu Jun 12 15:08:15 2025
  read: IOPS=10, BW=43.8MiB/s (45.9MB/s)(7880MiB/180073msec)
    slat (usec): min=54, max=7522, avg=149.70, stdev=313.45
    clat (msec): min=47, max=1900, avg=182.63, stdev=162.89
     lat (msec): min=48, max=1900, avg=182.78, stdev=162.89
    clat percentiles (msec):
     |  1.00th=[   80],  5.00th=[   90], 10.00th=[   97], 20.00th=[  109],
     | 30.00th=[  117], 40.00th=[  127], 50.00th=[  138], 60.00th=[  148],
     | 70.00th=[  165], 80.00th=[  199], 90.00th=[  296], 95.00th=[  435],
     | 99.00th=[ 1003], 99.50th=[ 1301], 99.90th=[ 1687], 99.95th=[ 1905],
     | 99.99th=[ 1905]
   bw (  KiB/s): min= 4724, max=81920, per=100.00%, avg=46489.33, stdev=20017.26, samples=345
   iops        : min=    1, max=   20, avg=10.94, stdev= 4.88, samples=345
  lat (msec)   : 50=0.05%, 100=12.39%, 250=74.42%, 500=9.24%, 750=2.08%
  lat (msec)   : 1000=0.86%, 2000=0.96%
  cpu          : usr=0.01%, sys=0.13%, ctx=2309, majf=0, minf=2059
  IO depths    : 1=0.1%, 2=99.9%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=1970,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=2

Run status group 0 (all jobs):
   READ: bw=43.8MiB/s (45.9MB/s), 43.8MiB/s-43.8MiB/s (45.9MB/s-45.9MB/s), io=7880MiB (8263MB), run=180073-180073msec
```

随机读:

```
fio  -directory=/mnt/3fs/ -numjobs=1 -fallocate=none -iodepth=2 -
ioengine=libaio -direct=1 -rw=randread -bs=4M --group_reporting -
size=100M -time_based -runtime=180 -name=2depth_128file_4M_direct_read_bw
```



```
root@prj2-vm1:~/MPI/MPI_Sort# fio  -directory=/mnt/3fs/ -numjobs=1 -fallocate=none -iodepth=2 -ioengine=libaio -direct=1 -rw=ran
dread -bs=4M --group_reporting -size=100M -time_based -runtime=180 -name=2depth_128file_4M_direct_read_bw
2depth_128file_4M_direct_read_bw: (g=0): rw=randread, bs=(R) 4096KiB-4096KiB, (W) 4096KiB-4096KiB, (T) 4096KiB-4096KiB, ioengine
=libaio, iodepth=2
fio-3.28
Starting 1 process
Jobs: 1 (f=1): [r(1)][100.0%][r=68.0MiB/s][r=17 IOPS][eta 00m:00s]
2depth_128file_4M_direct_read_bw: (groupid=0, jobs=1): err= 0: pid=7644: Thu Jun 12 15:17:41 2025
  read: IOPS=10, BW=40.4MiB/s (42.4MB/s)(7280MiB/180135msec)
    slat (usec): min=66, max=10933, avg=176.56, stdev=440.81
    clat (msec): min=68, max=1940, avg=197.71, stdev=177.94
     lat (msec): min=68, max=1940, avg=197.89, stdev=177.94
    clat percentiles (msec):
     |  1.00th=[   91],  5.00th=[  101], 10.00th=[  107], 20.00th=[  117],
     | 30.00th=[  127], 40.00th=[  136], 50.00th=[  146], 60.00th=[  161],
     | 70.00th=[  178], 80.00th=[  213], 90.00th=[  330], 95.00th=[  464],
     | 99.00th=[  986], 99.50th=[ 1418], 99.90th=[ 1938], 99.95th=[ 1938],
     | 99.99th=[ 1938]
   bw (  KiB/s): min= 3065, max=81920, per=100.00%, avg=43143.19, stdev=18905.90, samples=343
   iops        : min=    0, max=   20, avg=10.15, stdev= 4.62, samples=343
  lat (msec)   : 100=5.22%, 250=79.51%, 500=11.04%, 750=2.36%, 1000=0.93%
  lat (msec)   : 2000=0.93%
  cpu          : usr=0.02%, sys=0.13%, ctx=2165, majf=0, minf=2058
  IO depths    : 1=0.1%, 2=99.9%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=1820,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=2

Run status group 0 (all jobs):
   READ: bw=40.4MiB/s (42.4MB/s), 40.4MiB/s-40.4MiB/s (42.4MB/s-42.4MB/s), io=7280MiB (7634MB), run=180135-180135msec
```

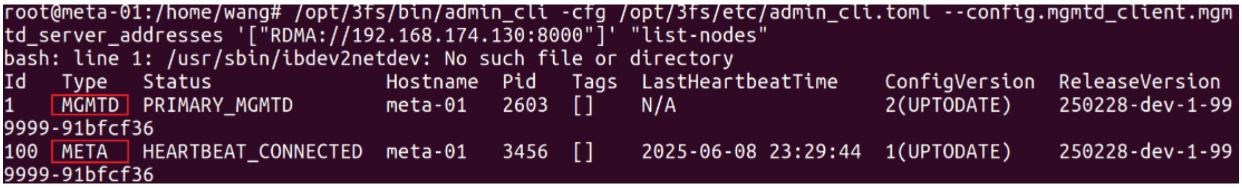
具体部署事项

写在前面

没有真实RDMA设备采用softRoCE方式时，外部必须是**Generic内核的完整Linux**环境，不管是编译部署还是docker部署，否则没有rxe模块。

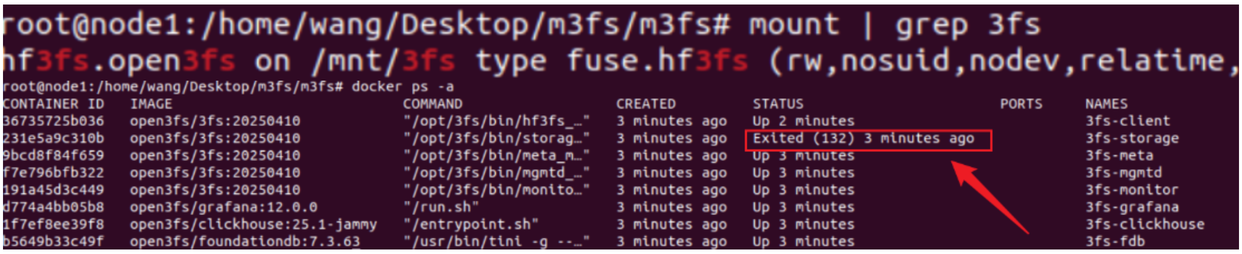
非常容易忽略的点，如果是硬件指令集没有avx512的时候，建议直接换有avx512机器。根据实践，使用虚拟机上是无法正常运行的（无论使用哪一种部署方式），哪怕给4G内存，128G的磁盘。在本地虚拟机上硬部署的结果（这个也有可能是本地的硬件带不动，但是由于之前服务器上的lxc容器是没有内核的，所以所有的尝试都在本地，大失败）：

1. 按照官方文件编译配置，编译成功，但是无法正常使用，尤其是Storage节点无法启动，直接卡死挂掉。



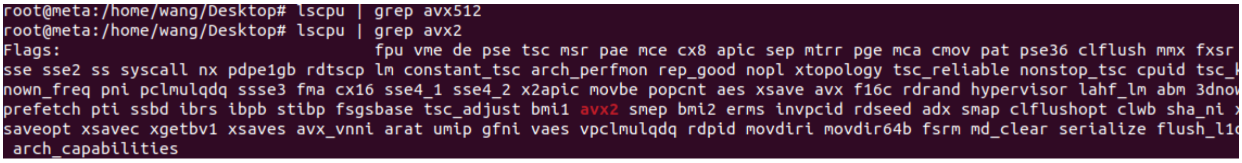
```
root@meta-01:/home/wang# /opt/3fs/bin/admin_cli -cfg /opt/3fs/etc/admin_cli.toml --config.mgmtd_client.mgm
td_server_addresses '["RDMA://192.168.174.130:8000"]' "list-nodes"
bash: line 1: /usr/sbin/ibdev2netdev: No such file or directory
Id   Type    Status               Hostname  Pid   Tags  LastHeartbeatTime    ConfigVersion  ReleaseVersion
1    MGMTD   PRIMARY_MGMTD        meta-01   2603  []    N/A                  2(UPTODATE)    250228-dev-1-99
9999-91bfcf36
100  META    HEARTBEAT_CONNECTED  meta-01   3456  []    2025-06-08 23:29:44  1(UPTODATE)    250228-dev-1-99
9999-91bfcf36
```

2. 按照docker方式部署，部署成功，挂载成功，但是storage异常退出，退出码351，为指令不兼容。使用-avx2版本的结果：通官方编译结果，直接卡死，挂掉，而且难以收集异常退出信息。



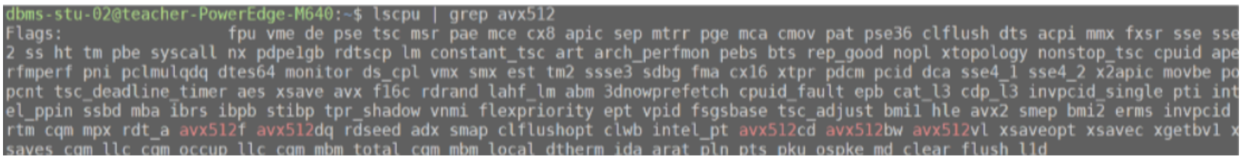
```
root@node1:/home/wang/Desktop/m3fs/m3fs# mount | grep 3fs
hf3fs.open3fs on /mnt/3fs type fuse.hf3fs (rw,nosuid,nodev,relatime,
root@node1:/home/wang/Desktop/m3fs/m3fs# docker ps -a
CONTAINER ID   IMAGE                             COMMAND                  CREATED         STATUS                       PORTS     NAMES
36735725b036   open3fs/3fs:20250410              "/opt/3fs/bin/hf3fs_…"   3 minutes ago   Up 2 minutes                           3fs-client
231e5a9c310b   open3fs/3fs:20250410              "/opt/3fs/bin/storag…"   3 minutes ago   Exited (132) 3 minutes ago             3fs-storage
9bcd8f84f659   open3fs/3fs:20250410              "/opt/3fs/bin/meta_m…"   3 minutes ago   Up 3 minutes                           3fs-meta
f7e796bfb322   open3fs/3fs:20250410              "/opt/3fs/bin/mgmtd_…"   3 minutes ago   Up 3 minutes                           3fs-mgmtd
191a45d3c449   open3fs/3fs:20250410              "/opt/3fs/bin/monito…"   3 minutes ago   Up 3 minutes                           3fs-monitor
d774a4bb05b8   open3fs/grafana:12.0.0            "/run.sh"                3 minutes ago   Up 3 minutes                           3fs-grafana
1f7ef8ee39f8   open3fs/clickhouse:25.1-jammy     "/entrypoint.sh"         3 minutes ago   Up 3 minutes                           3fs-clickhouse
b5649b33c49f   open3fs/foundationdb:7.3.63       "/usr/bin/tini -g --…"   3 minutes ago   Up 3 minutes                           3fs-fdb
```

完全没成功的lscpu，只有avx2：



```
root@meta:/home/wang/Desktop# lscpu | grep avx512
root@meta:/home/wang/Desktop# lscpu | grep avx2
Flags:                              fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr
sse sse2 ss syscall nx pdpe1gb rdtscp lm constant_tsc arch_perfmon rep_good nopl xtopology tsc_reliable nonstop_tsc cpuid tsc_k
nown_freq pni pclmulqdq ssse3 fma cx16 sse4_1 sse4_2 x2apic movbe popcnt aes xsave avx f16c rdrand hypervisor lahf_lm abm 3dnow
prefetch pti ssbd ibrs ibpb stibp fsgsbase tsc_adjust bmi1 avx2 smep bmi2 erms invpcid rdseed adx smap clflushopt clwb sha_ni x
saveopt xsavec xgetbv1 xsaves avx_vnni arat umip gfni vaes vpclmulqdq rdpid movdiri movdir64b fsrm md_clear serialize flush_l1d
 arch_capabilities
```

成功的lscpu，有avx512：

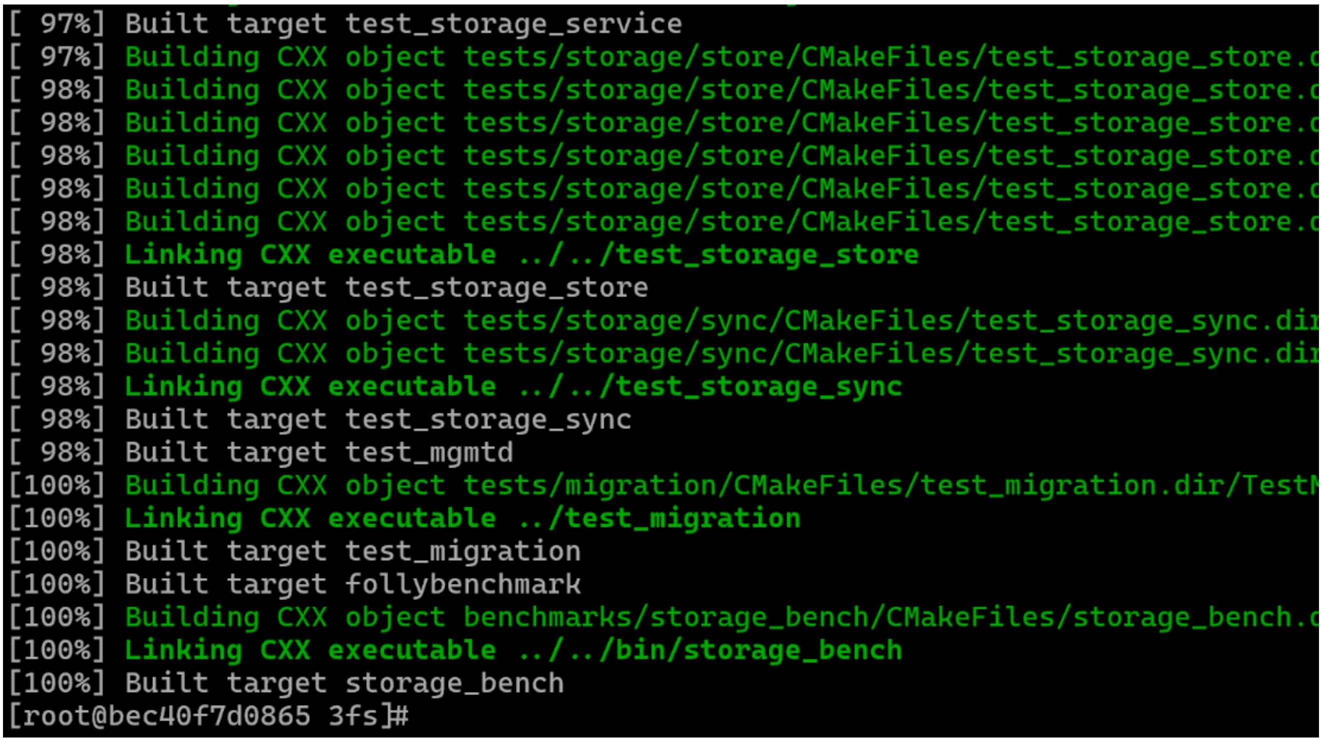


```
dbms-stu-02@teacher-PowerEdge-M640:~$ lscpu | grep avx512
Flags:                  fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse
2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm constant_tsc art arch_perfmon pebs bts rep_good nopl xtopology nonstop_tsc cpuid ape
rfmperf pni pclmulqdq dtes64 monitor ds_cpl vmx smx est tm2 ssse3 sdbg fma cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic movbe po
pcnt tsc_deadline_timer aes xsave avx f16c rdrand lahf_lm abm 3dnowprefetch cpuid_fault epb cat_l3 cdp_l3 invpcid_single pti int
el_ppin ssbd mba ibrs ibpb stibp tpr_shadow vnmi flexpriority ept vpid fsgsbase tsc_adjust bmi1 hle avx2 smep bmi2 erms invpcid
rtm cqm mpx rdt_a avx512f avx512dq rdseed adx smap clflushopt clwb intel_pt avx512cd avx512bw avx512vl xsaveopt xsavec xgetbv1 x
saves cqm_llc cqm_occup_llc cqm_mbm_total cqm_mbm_local dtherm ida arat pln pts pku ospke md_clear flush_l1d
```

部署方式一：编译文件

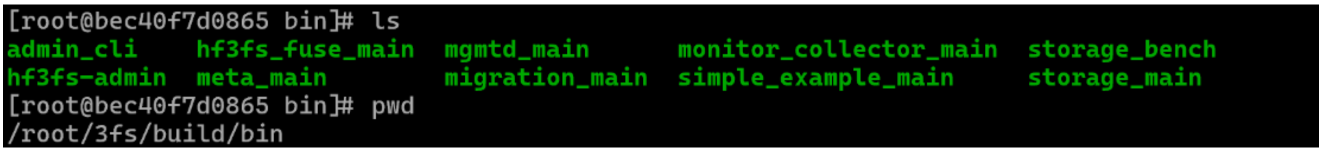
按照官方文件进行，编译安装。问题是编译安装时间过长，而且配置文件的修改需要自己手动修改，有些冗余。好处是知道各种结构，搭配配置文件可塑性强。

编译结果：



```
[ 97%] Built target test_storage_service
[ 97%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Building CXX object tests/storage/store/CMakeFiles/test_storage_store.d
[ 98%] Linking CXX executable ../../test_storage_store
[ 98%] Built target test_storage_store
[ 98%] Building CXX object tests/storage/sync/CMakeFiles/test_storage_sync.dir
[ 98%] Building CXX object tests/storage/sync/CMakeFiles/test_storage_sync.dir
[ 98%] Linking CXX executable ../../test_storage_sync
[ 98%] Built target test_storage_sync
[ 98%] Built target test_mgmtd
[100%] Building CXX object tests/migration/CMakeFiles/test_migration.dir/TestM
[100%] Linking CXX executable ../test_migration
[100%] Built target test_migration
[100%] Built target follybenchmark
[100%] Building CXX object benchmarks/storage_bench/CMakeFiles/storage_bench.d
[100%] Linking CXX executable ../../bin/storage_bench
[100%] Built target storage_bench
[root@bec40f7d0865 3fs]#
```

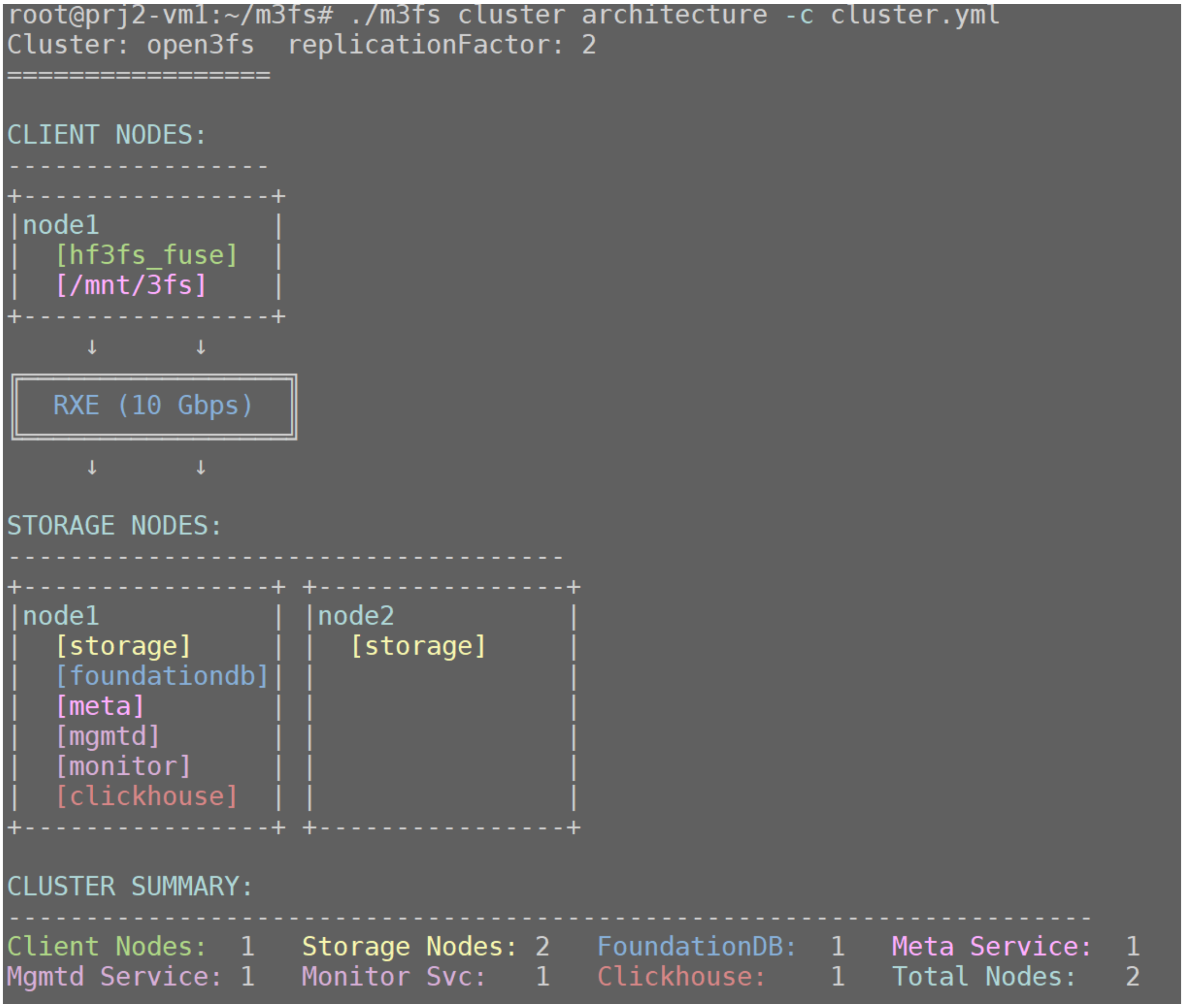
可执行文件：



```
[root@bec40f7d0865 bin]# ls
admin_cli     hf3fs_fuse_main  mgmtd_main      monitor_collector_main  storage_bench
hf3fs-admin   meta_main        migration_main  simple_example_main     storage_main
[root@bec40f7d0865 bin]# pwd
/root/3fs/build/bin
```

部署方式二：docker部署

集群结构：

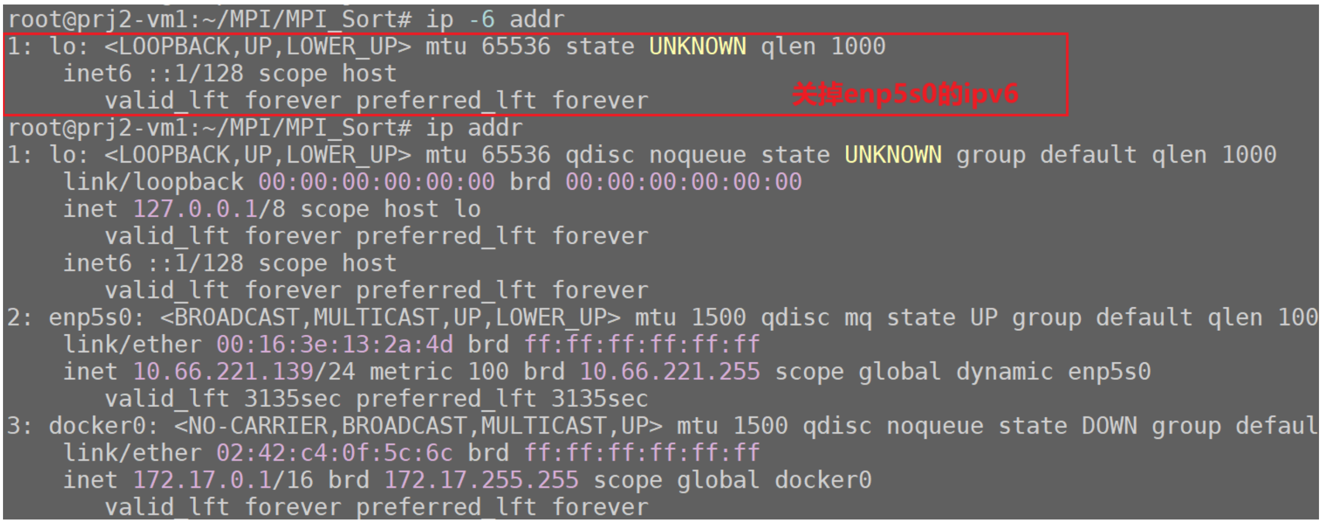


```
root@prj2-vm1:~/m3fs# ./m3fs cluster architecture -c cluster.yml
Cluster: open3fs  replicationFactor: 2
=================

CLIENT NODES:
-----------------
+-----------------+
|node1            |
|  [hf3fs_fuse]   |
|  [/mnt/3fs]     |
+-----------------+
     ↓      ↓
╔══════════════════╗
║  RXE (10 Gbps)   ║
╚══════════════════╝
     ↓      ↓

STORAGE NODES:
------------------------------------
+-----------------+ +-----------------+
|node1            | |node2            |
|  [storage]      | |  [storage]      |
|  [foundationdb]| |                 |
|  [meta]         | |                 |
|  [mgmtd]        | |                 |
|  [monitor]      | |                 |
|  [clickhouse]   | |                 |
+-----------------+ +-----------------+

CLUSTER SUMMARY:
-----------------------------------------------------------------------
Client Nodes:   1   Storage Nodes: 2   FoundationDB:   1   Meta Service:   1
Mgmtd Service: 1   Monitor Svc:    1   Clickhouse:     1   Total Nodes:    2
```

来自：https://github.com/open3fs/m3fs

这里的主要问题就是在lxc的Ubuntu容器里对网络的要求有些奇怪：foundationDB的docker中需要ipv4，否则挂掉（无法解析）；同时clickhouse的docker中又要求不能完全直接关闭ipv6。但是在直接的Ubuntu虚拟机中没有这些奇怪的毛病。

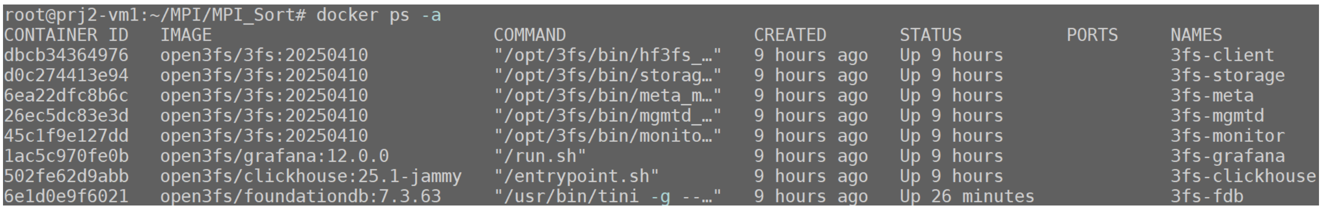
其他的就是保证端口没有被占用即可。



```
root@prj2-vm1:~/MPI/MPI_Sort# ip -6 addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 state UNKNOWN qlen 1000
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
root@prj2-vm1:~/MPI/MPI_Sort# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp5s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen 100
    link/ether 00:16:3e:13:2a:4d brd ff:ff:ff:ff:ff:ff
    inet 10.66.221.139/24 metric 100 brd 10.66.221.255 scope global dynamic enp5s0
       valid_lft 3135sec preferred_lft 3135sec
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN group defaul
    link/ether 02:42:c4:0f:5c:6c brd ff:ff:ff:ff:ff:ff
    inet 172.17.0.1/16 brd 172.17.255.255 scope global docker0
       valid_lft forever preferred_lft forever
```

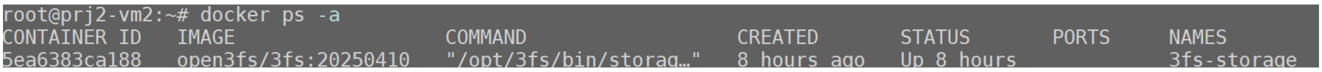
关掉enp5s0的ipv6

部署成功界面



```
root@prj2-vm1:~/MPI/MPI_Sort# docker ps -a
CONTAINER ID   IMAGE                          COMMAND                  CREATED       STATUS          PORTS     NAMES
dbcb34364976   open3fs/3fs:20250410           "/opt/3fs/bin/hf3fs_…"   9 hours ago   Up 9 hours                3fs-client
d0c274413e94   open3fs/3fs:20250410           "/opt/3fs/bin/storag…"   9 hours ago   Up 9 hours                3fs-storage
6ea22dfc8b6c   open3fs/3fs:20250410           "/opt/3fs/bin/meta_m…"   9 hours ago   Up 9 hours                3fs-meta
26ec5dc83e3d   open3fs/3fs:20250410           "/opt/3fs/bin/mgmtd_…"   9 hours ago   Up 9 hours                3fs-mgmtd
45c1f9e127dd   open3fs/3fs:20250410           "/opt/3fs/bin/monito…"   9 hours ago   Up 9 hours                3fs-monitor
1ac5c970fe0b   open3fs/grafana:12.0.0         "/run.sh"                9 hours ago   Up 9 hours                3fs-grafana
502fe62d9abb   open3fs/clickhouse:25.1-jammy  "/entrypoint.sh"         9 hours ago   Up 9 hours                3fs-clickhouse
6e1d0e9f6021   open3fs/foundationdb:7.3.63    "/usr/bin/tini -g --…"   9 hours ago   Up 26 minutes             3fs-fdb
```

另一个storage节点的部署：

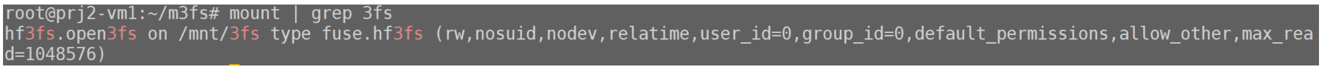


```
root@prj2-vm2:~# docker ps -a
CONTAINER ID   IMAGE                  COMMAND                  CREATED       STATUS       PORTS     NAMES
5ea6383ca188   open3fs/3fs:20250410   "/opt/3fs/bin/storag…"   8 hours ago   Up 8 hours             3fs-storage
```

挂载目录

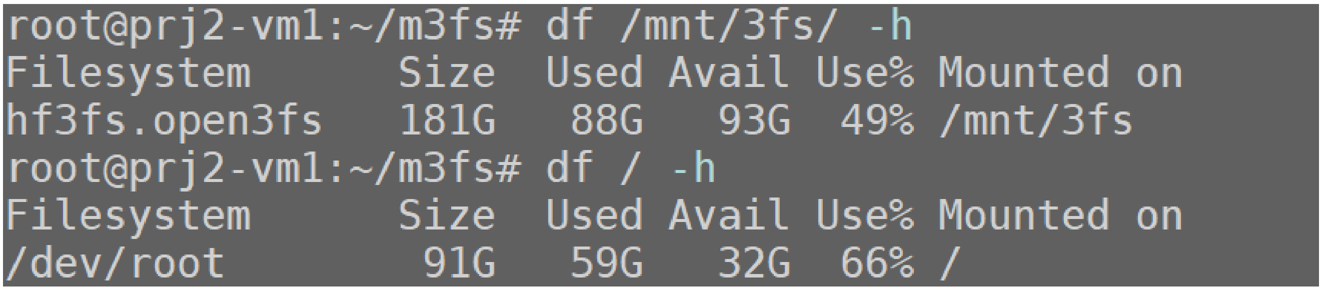
```
mount | grep 3fs
```

读写方式就和直接读写本地文件系统一样【fuse的方式，便捷，性能会差点：https://github.com/deepseek-ai/3FS/blob/main/docs/design_notes.md#limitations-of-fuse】。当然还有别的方式：https://github.com/deepseek-ai/3FS/blob/main/docs/design_notes.md#asynchronous-zero-copy-api



```
root@prj2-vm1:~/m3fs# mount | grep 3fs
hf3fs.open3fs on /mnt/3fs type fuse.hf3fs (rw,nosuid,nodev,relatime,user_id=0,group_id=0,default_permissions,allow_other,max_read=1048576)
```

挂载大小以及FileSystem：



```
root@prj2-vm1:~/m3fs# df /mnt/3fs/ -h
Filesystem       Size  Used Avail Use% Mounted on
hf3fs.open3fs    181G   88G   93G  49% /mnt/3fs
root@prj2-vm1:~/m3fs# df / -h
Filesystem       Size  Used Avail Use% Mounted on
/dev/root         91G   59G   32G  66% /
```