IBM

IBM Spectrum Scale

IBM SpectrumScale – UG ISC 2022, update

olaf.weiser@de.ibm.com

Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

Agenda

Resource planning for MCOT

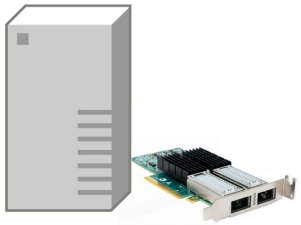
Field report I – bond
Field report II – file creation

Performance – new ESS3500

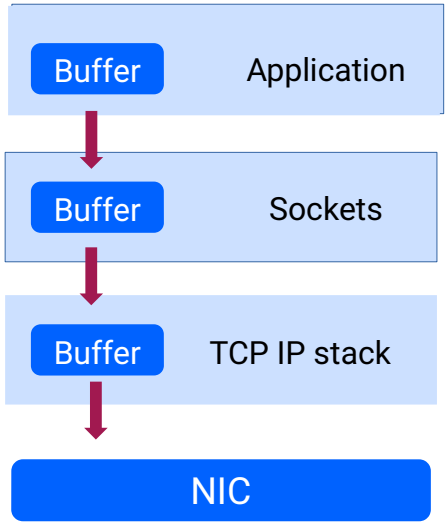
Research project

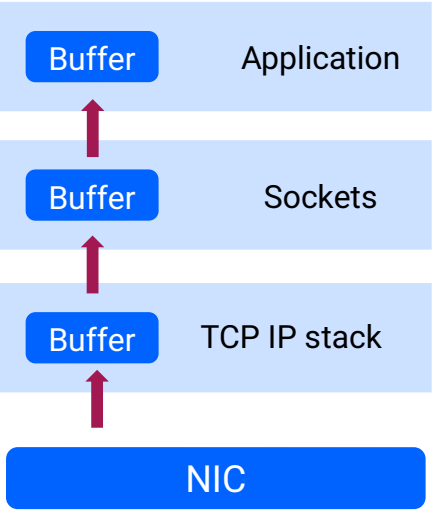
MultipleConnectionTCP - MCOT

IBM ESS 3500

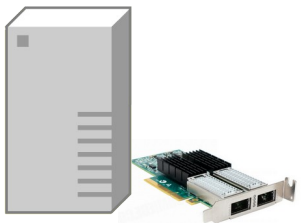


100 EDR or 200 HDR (400 NDR)

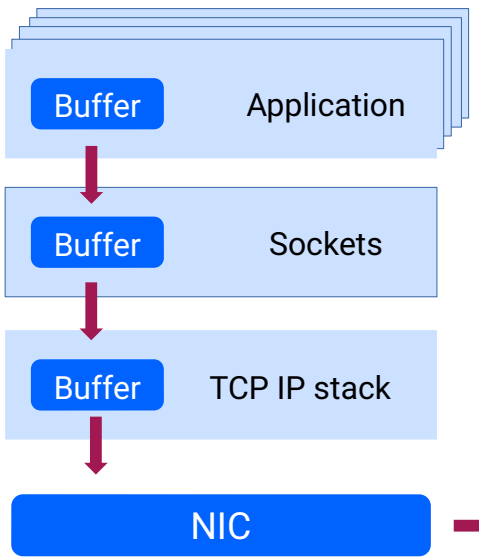


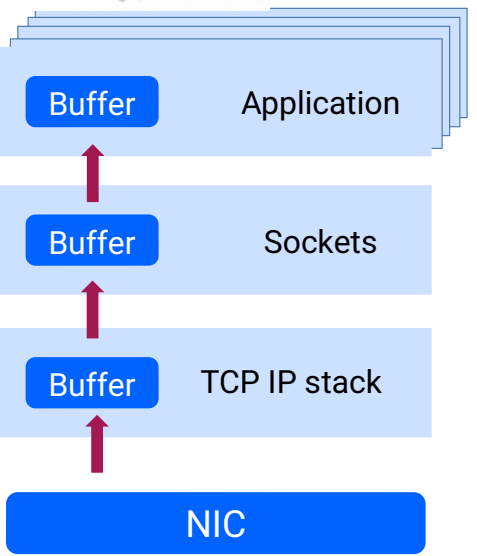


MultipleConnectionTCP - MCOT

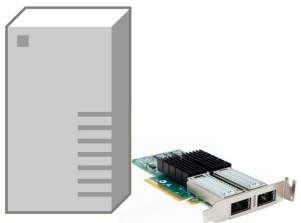


100 EDR or 200 HDR (400 NDR)

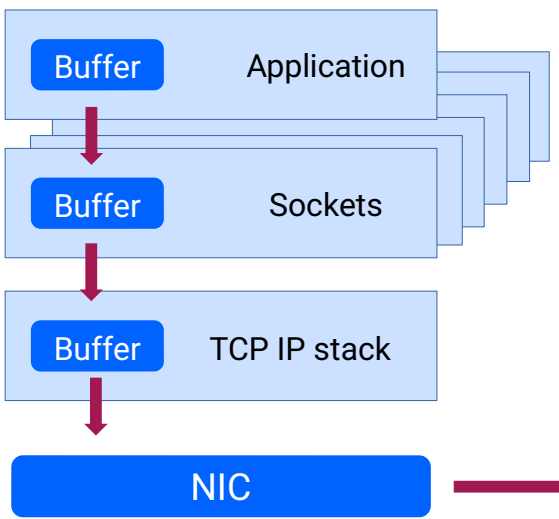




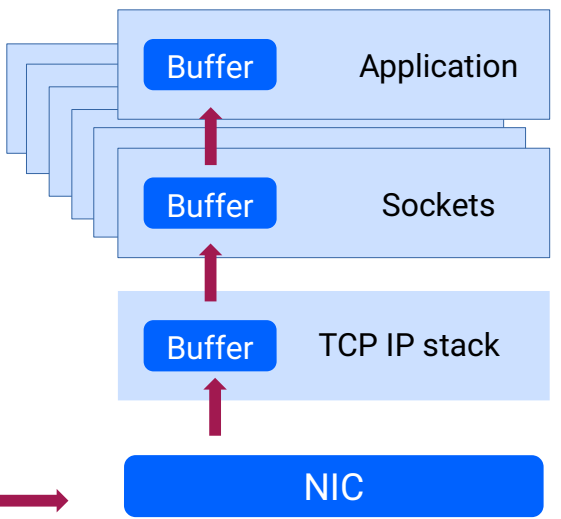
MultipleConnectionTCP - MCOT



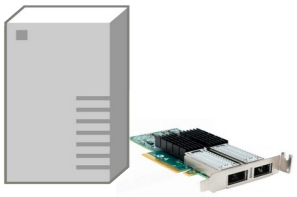
100 EDR or 200 HDR (400 NDR)

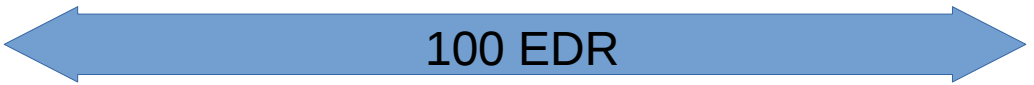


Using multiple sockets
generate some scaling effects



Single socket - iperf





```
[root@fscc-sr650-54 beer8m]# iperf3 -c  10.200.1.3 -w 1M
Connecting to host 10.200.1.3, port 5201
[  5] local 10.200.1.7 port 44750 connected to 10.200.1.3 port 5201
[ ID] Interval           Transfer     Bitrate         Retr  Cwnd
[  5]   0.00-1.00   sec  4.86 GBytes  41.7 Gbits/sec    0   2.77 MBytes
[  5]   1.00-2.00   sec  4.67 GBytes  40.1 Gbits/sec    0   2.77 MBytes
[  5]   2.00-3.00   sec  4.67 GBytes  40.1 Gbits/sec    0   2.77 MBytes
[  5]   3.00-4.00   sec  4.62 GBytes  39.6 Gbits/sec    0   2.77 MBytes
[  5]   4.00-5.00   sec  4.77 GBytes  40.9 Gbits/sec    0   2.77 MBytes
[  5]   5.00-6.00   sec  4.72 GBytes  40.6 Gbits/sec    0   2.77 MBytes
[  5]   6.00-7.00   sec  4.76 GBytes  40.9 Gbits/sec    0   2.77 MBytes
[  5]   7.00-8.00   sec  4.75 GBytes  40.8 Gbits/sec    0   2.77 MBytes
[  5]   8.00-9.00   sec  4.74 GBytes  40.8 Gbits/sec    0   2.77 MBytes
[  5]   9.00-10.00  sec  4.75 GBytes  40.8 Gbits/sec    0   2.77 MBytes
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bitrate         Retr
[  5]   0.00-10.00  sec  47.3 GBytes  40.6 Gbits/sec    0             sender
[  5]   0.00-10.04  sec  47.3 GBytes  40.5 Gbits/sec                  receiver
```

~ 5GB/s

maxTcpConnsPerNodeConn = 1 (single socket)

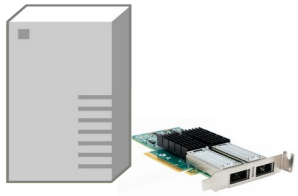
maxTcpConnsPerNodeConn=1

100 EDR

[root@fscc-sr650-54 beer8m]# time /usr/lpp/mmfs/samples/perf/gpfsperf create seq myfilefoo1 -n 200g -r 8m -th 12
/usr/lpp/mmfs/samples/perf/gpfsperf create seq myfilefoo1
  recSize 8M nBytes 200G fileSize 200G
  nProcesses 1 nThreadsPerProcess 12
  file cache flushed before test
  not using direct I/O
  offsets accessed will cycle through the same file segment
  not using shared memory buffer
  not releasing byte-range token after open
  no fsync at end of test
  Data rate was 5089216.97 Kbytes/sec, Op Rate was 606.68 Ops/sec, Avg Latency was 19.350 milliseconds, thread utilization 0.978, bytesTransf

```
real    0m42.281s
user    0m0.079s
sys     1m5.276s
```
[root@fscc-sr650-54 beer8m]#

maxTcpConnsPerNodeConn = 1 (single socket)



maxTcpConnsPerNodeConn=1

100 EDR

Data rate was 5089216.97 Kbytes/sec, Op Rate was 606.68 Ops/sec, Avg Latency was 19.350 milliseconds, thread utilization 0.978, bytesTransf

```
real    0m42.281s
user    0m0.079s
sys     1m5.276s
[root@fscc-sr650-54 beer8m]#
```

atop:

```
CPU | sys     293% | user     28% | irq      18%          | 151%          gpfsperf
CPU | sys     295% | user     84% | irq      45%          | 162%          gpfsperf
CPU | sys     287% | user     83% | irq      47%          | 165%          gpfsperf
CPU | sys     280% | user     83% | irq      44%          | 164%          gpfsperf
CPU | sys     295% | user     84% | irq      47%
```

maxTcpConnsPerNodeConn = 2

maxTcpConnsPerNodeConn=2

100 EDR

Data rate was 7977541.20 Kbytes/sec, Op Rate was 606.68 Ops/sec, Avg Latency was 19.350 milliseconds, thread utilization 0.978, bytesTransf

```
real    0m27.002s
user    0m0.085s
sys     1m17.176s
```

atop:

```
CPU | sys     375% | user    133% | irq      63%
CPU | sys     530% | user    178% | irq      80%
CPU | sys     396% | user    115% | irq      62%
CPU | sys     568% | user    174% | irq      83%
CPU | sys     460% | user    150% | irq      71%
```

```
| 280%          gpfsperf
| 291%          gpfsperf
| 252%          gpfsperf
| 267%          gpfsperf
```

maxTcpConnsPerNodeConn = 4

maxTcpConnsPerNodeConn=4

100 EDR

Data rate was 8416153.42 Kbytes/sec, Op Rate was 606.68 Ops/sec, Avg Latency was 19.350 milliseconds, thread utilization 0.978, bytesTransf

```
real    0m25.600s
user    0m0.080s
sys     1m17.877s
[root@fscc-sr650-54 beer8m]#
```

atop:

```
CPU | sys     597% | user    189% | irq      92%
CPU | sys     795% | user    324% | irq     109%
CPU | sys     832% | user    314% | irq     112%
CPU | sys     634% | user    217% | irq      99%
```

```
267%          gpfsperf
334%          gpfsperf
376%          gpfsperf
376%          gpfsperf
```

maxTcpConnsPerNodeConn = 8

maxTcpConnsPerNodeConn=8

100 EDR

Data rate was 9059055.19 Kbytes/sec, Op Rate was 606.68 Ops/sec, Avg Latency was 19.350 milliseconds, thread utilization 0.978, bytesTransf

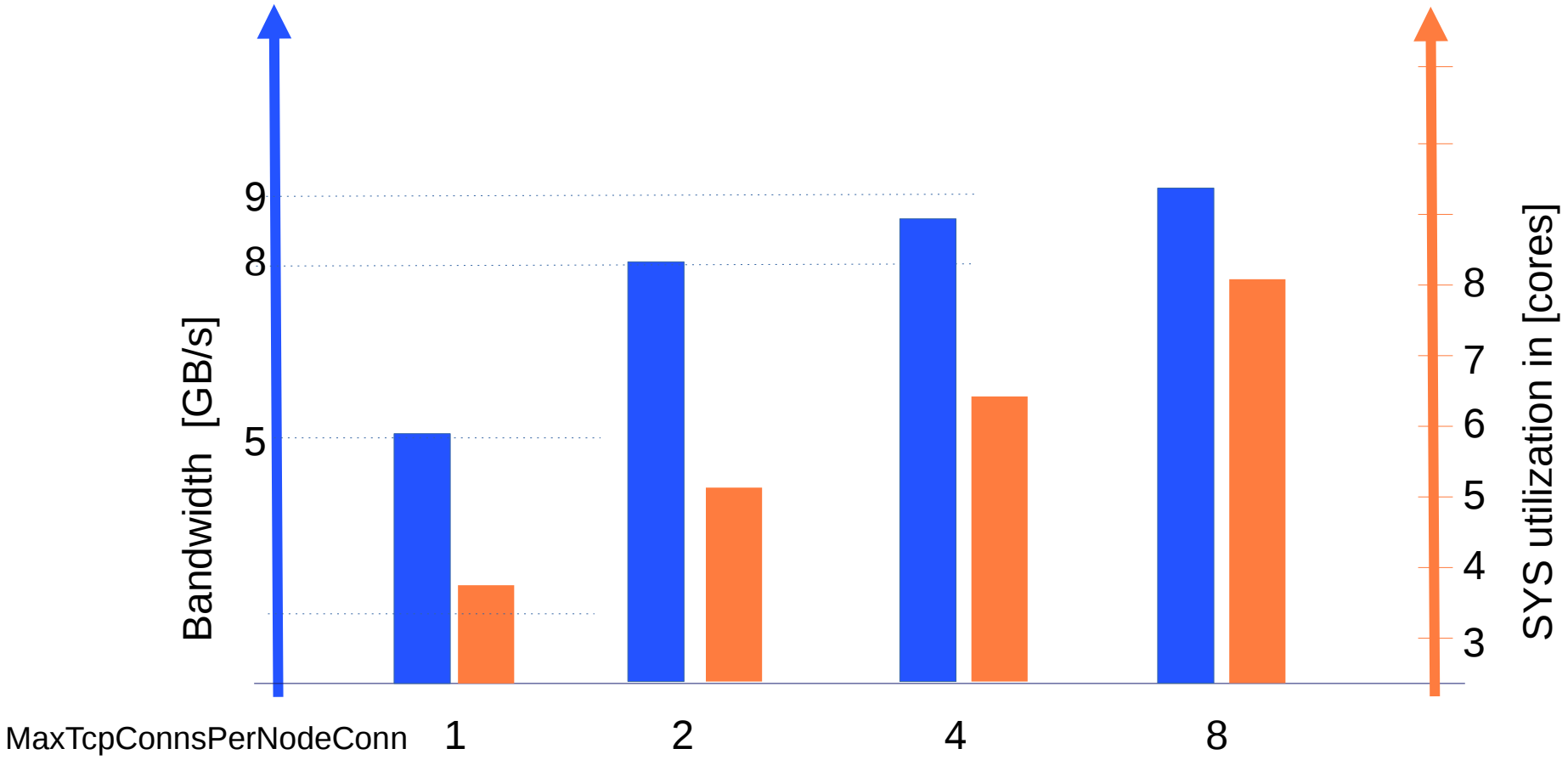
```
real    0m23.787s
user    0m0.077s
sys     1m27.027s
[root@fscc-sr650-54 beer8m]#
```

atop:

```
CPU | sys     794% | user    228% | irq     114%
CPU | sys     731% | user    216% | irq     111%
CPU | sys     720% | user    224% | irq     117%
CPU | sys     810% | user    207% | irq      99%
```

```
395%          gpfsperf
334%          gpfsperf
396%          gpfsperf
364%          gpfsperf
```

Core / bandwidth ratio using TCPIP



RDMA / RoCE

100 EDR

Data rate was 11276356.48 Kbytes/sec, Op Rate was 1344.25 Ops/sec, Avg Latency was 8.760 milliseconds, …

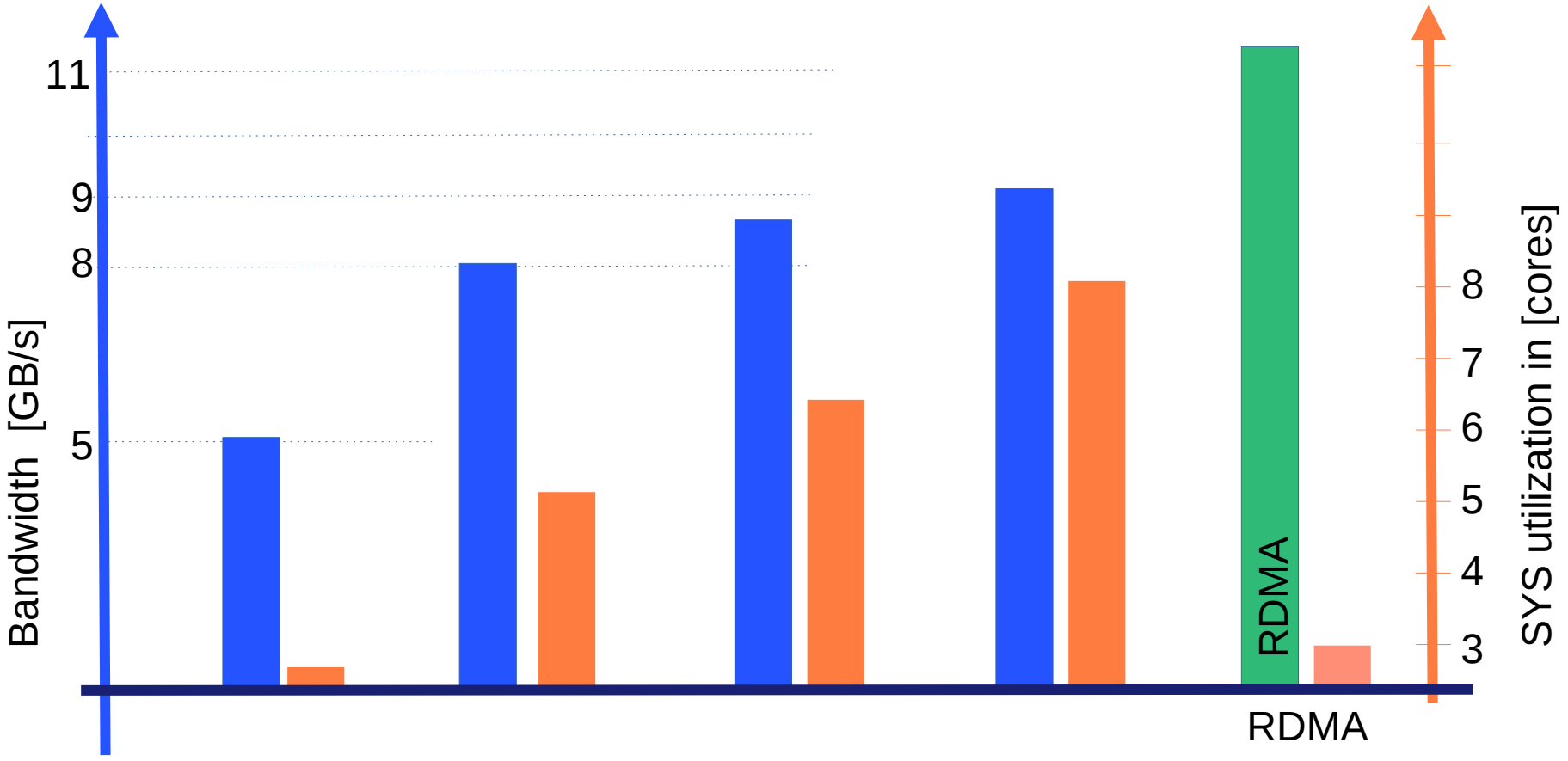
```
real    0m19.126s
user    0m0.076s
sys     0m58.793s


[root@fscc-sr650-54 beer8m]#
```

atop:

```
CPU | sys     340%    user    258% | irq       9%
CPU | sys     356%    user    256% | irq       8%
CPU | sys     361%    user    237% | irq       7%
CPU | sys     244%    user    191% | irq       7%
```

```
330%          gpfsperf
320%          gpfsperf
313%          gpfsperf
306%          gpfsperf
```

Core / bandwidth ratio using TCPIP



Agenda

Field report I – bond
Field report II – file creation

Performance – new ESS3500

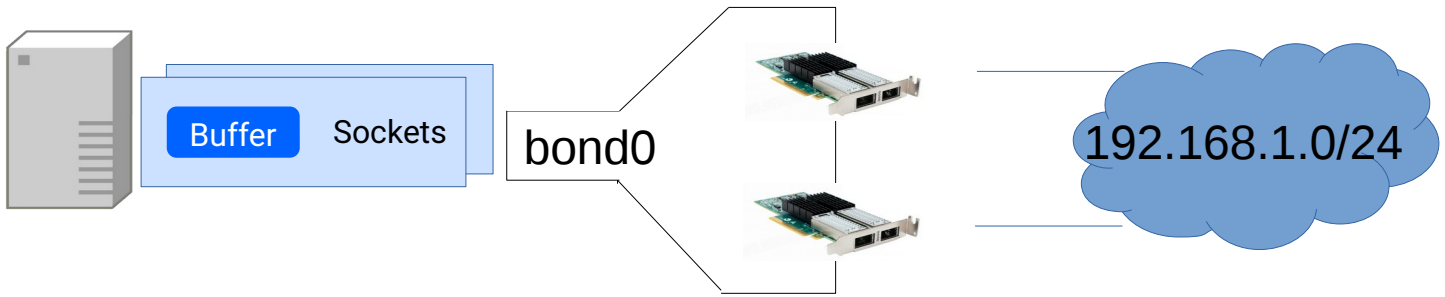
Research project

TCP/IP communication with…

… MCOT
… mulitple (subnets)~ networks

[M]LAG , maxTcpConnsPerNodeConn and bond configuration

– MCOT helps to utilize more than one network port also for TCP/IP
– keep in mind, the CPU(core) utilization / SYS load for TCP



– to benefit from MCOT in a flat network with LAG (bonds) , use

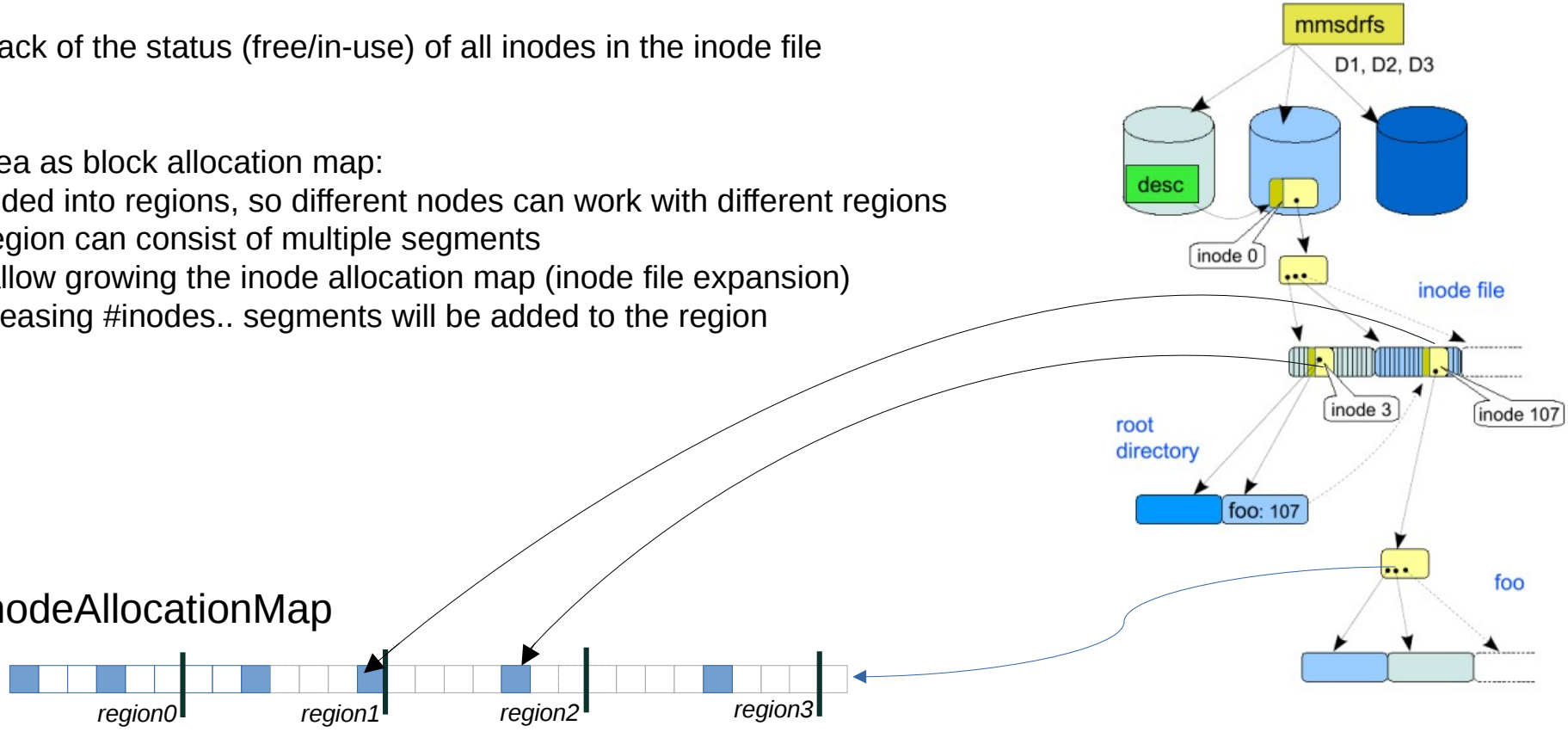
```
xmit_hash_policy=layer3+4
```

File creation workloads

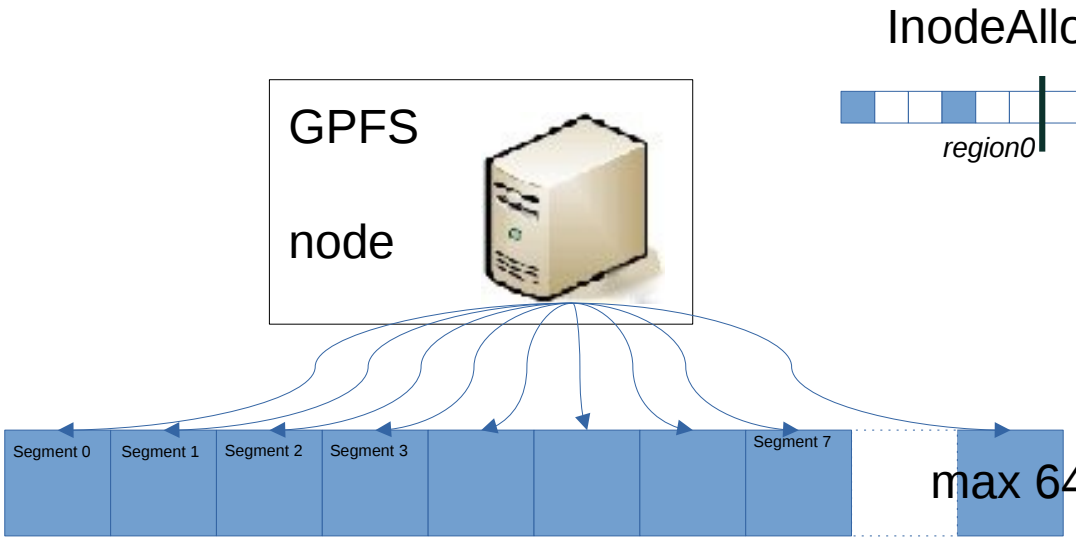
Inode Allocation Map

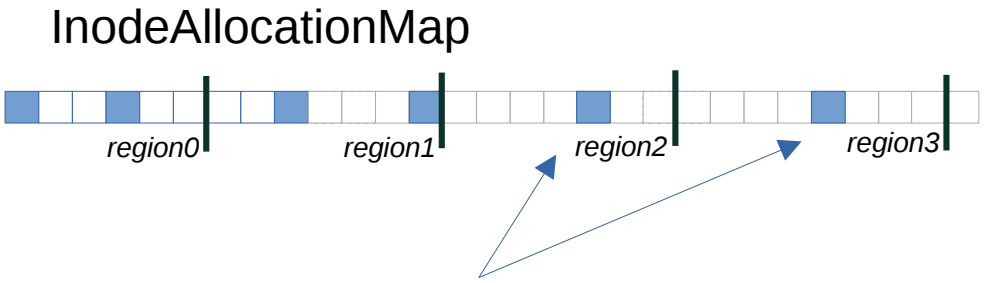
- Keeps track of the status (free/in-use) of all inodes in the inode file

- Same idea as block allocation map:
  - Divided into regions, so different nodes can work with different regions
  - A region can consist of multiple segments
    to allow growing the inode allocation map (inode file expansion)
  - increasing #inodes.. segments will be added to the region



Active Inode Allocation regions and segments





Inode allocation segments

Default: -1
Range: 1-64
File system format version: 5.0.2 or later

mmchconfig maxActiveIallocSegs=default (which is 8 or 4 , depeding on -n)

Specifies the number of active inode allocation segments that are maintained on the specified nodes. The valid range is 1 - 64. A value greater than 1 can significantly improve performance in the following scenario:

A single node has created a large number of files in multiple directories.

Processes and threads on multiple nodes are now concurrently attempting to delete or unlink files in those directories.

Active Inode Allocation regions and segments

– number of regions depends on #disk and -n value from mmcrfs

– ‘-n’ value can be changed for existing file system but it has no effect
for the inode allocation map

– shortly after creating the file system, you’ll get 8x or 4x the number of ‘-n’ regions

– inodes retrieved then from segments,

– design/plan accordingly, that enough regions are available (‘-n’)

– the more inodes are used/needed, the more segments will be needed (done automatically)

– segments are distributed equally across all regions

Ad hoc improvement: Avoid long delay for finding new inodes

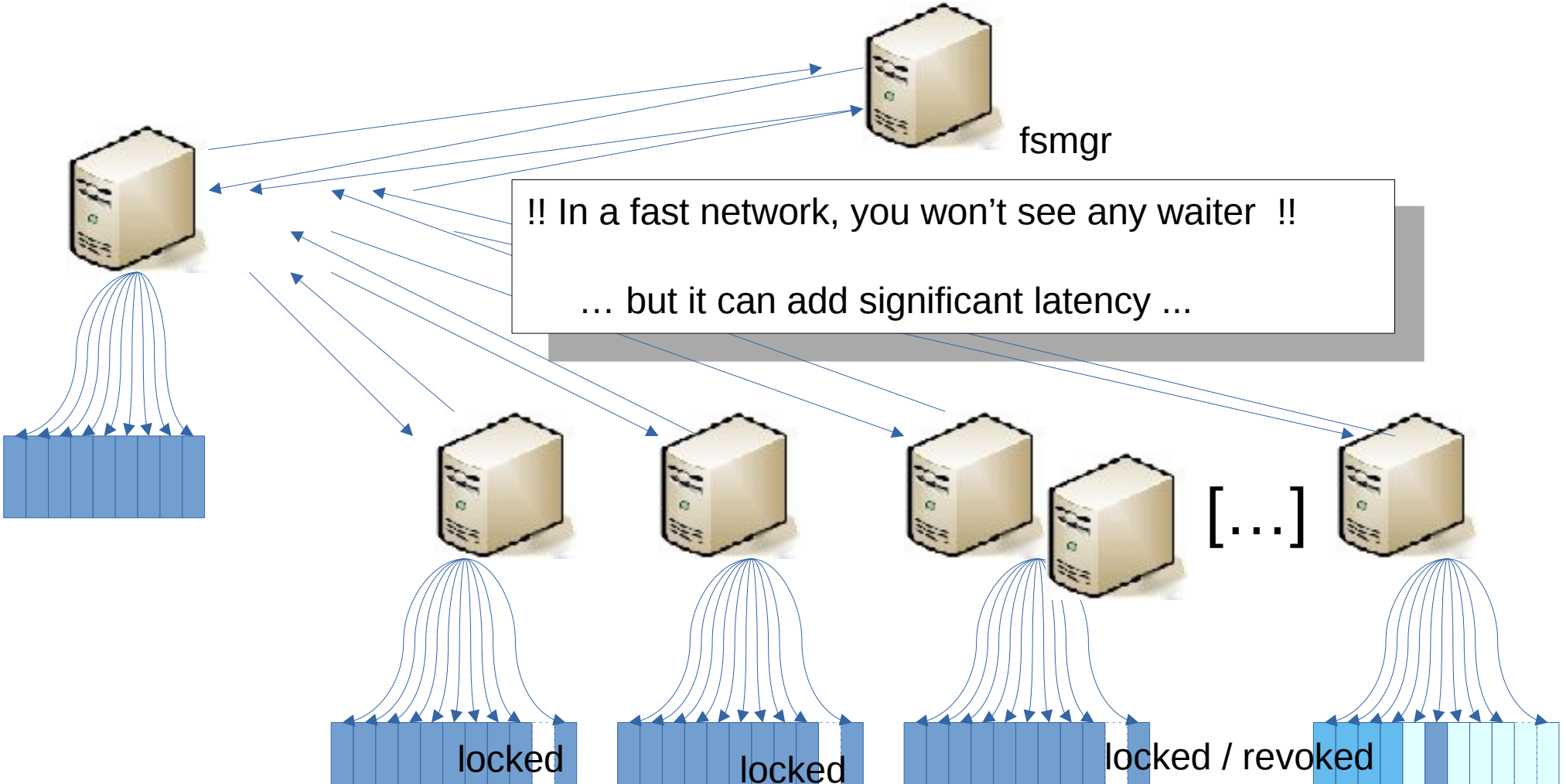
[root@asap3-utl04 ~]# mmdf core1 -F -q
[…]

```
Inode Information
-----------------
Total number of used inodes in all Inode spaces:          613765384
Total number of free inodes in all Inode spaces:          10822904
Total number of allocated inodes in all Inode spaces:     624588288
Total of Maximum number of inodes in all Inode spaces:    794424832
```

```
[root@asap3-utl04 ~]# mmlsmount
File system rest is mounted on 827 nodes.
[root@asap3-utl04 ~]# mmlsfs […] -n
 -n                 2000                       Estimated number of nodes that will \
                                               mount file system
```

10822904 / 2000 =5400 Inodes (statistically)

.. It may generate a lot of network traffic..



A closer look behind the scenes

```
36.200570261  27127 TRACE_IALLOC: accessIAlloc: ireg 654 iseg 32 lock failed, err 16.
36.201321722  27127 TRACE_IALLOC: accessIAlloc: ireg 2618 iseg 32 lock failed, err 16

...

36.763190624  27127 TRACE_IALLOC: accessIAlloc: ireg 6311 iseg 32 lock failed, err 16
36.763190932  27127 TRACE_IALLOC: allocInode: phase 2 unsuccessful
37.222842931  27127 TRACE_IALLOC: accessIAlloc: ireg 2842 iseg 0 lock failed, err 16
37.326247940  27127 TRACE_IALLOC: accessIAlloc: ireg 3195 iseg 0 lock failed, err 16

...

39.265700074  27127 TRACE_IALLOC: accessIAlloc: ireg 2237 iseg 8 lock failed, err 16
39.367316134  27127 TRACE_IALLOC: accessIAlloc: ireg 3790 iseg 8 lock failed, err 16
39.886688617  27127 TRACE_IALLOC: accessIAlloc: ireg 2715 iseg 8 lock failed, err 16
```

Ad-hoc solution: Provide more free/preallocated inodes

```
[root@ess5kio1 ~]# mmdf ess5k8m
[…]
Inode Information
-----------------
Total number of used inodes in all Inode spaces:          144968905
Total number of free inodes in all Inode spaces:           73335607
Total number of allocated inodes in all Inode spaces:     218304512
Total of Maximum number of inodes in all Inode spaces:    424312832


[root@ess5kio1 ~]# time mmchfs ess5k8m --inode-limit 1342377984:424110208
Set maxInodes for inode space 0 to 1342377984
Fileset root changed.

real    7m37.331s
user    0m0.649s
sys     0m0.081s
[root@ess5kio1 ~]#
```

… be patient, it can take a while …

Single Node, 16 threads (cores) mdtest

mdtest-3.3.0 was launched with 16 total task(s) on 1 node(s)

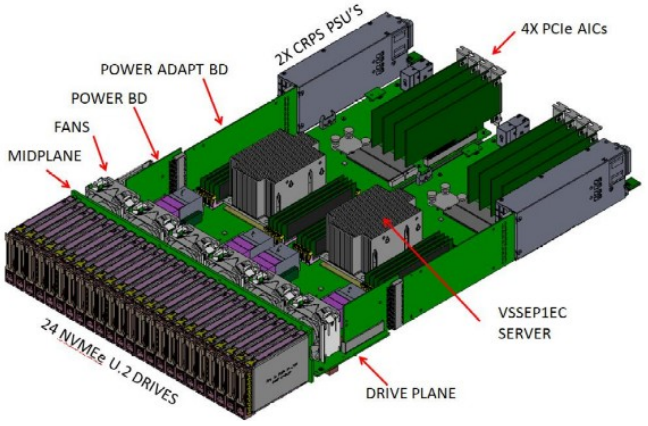
Command line used: mdtest '-I' '800' '-i' '1' '-u' '-t' '-z' '4' '-b' '3' '-p' '15' '-y' '-v' '-d=/gpfs/gpfs0/benchmark/rhel82/data/'

```
SUMMARY rate: (of 1 iterations)
   Operation                     Max            Min           Mean        Std Dev
   ---------                     ---            ---           ----        -------
   Directory creation        :      25182.231      25181.256      25182.043          0.303
   Directory stat            :     413217.983     412954.577     413005.793         81.750
   Directory removal         :      32592.325      32592.317      32592.323          0.002
   File creation             :      28790.083      28790.061      28790.075          0.008
   File stat                 :     501834.631     501830.371     501832.607          0.857
   File read                 :     363917.795     363916.511     363916.763          0.310
   File removal              :      43487.590      43487.586      43487.587          0.001
   Tree creation             :       1659.101       1659.101       1659.101          0.000
   Tree removal              :        119.390        119.390        119.390          0.000
V-1: Entering PrintTimestamp...
-- finished at 11/10/2021 13:00:09 --
```

The new ESS3500

IBM ESS 3500

ESS 3500 – some numbers from next generation ESS



PCI Layout :

– 128 Lanes Gen4 16GT/s
– x2 x 24 to Disks
– x16 x 4 for Adapters

IBM ESS 3500

READ:	up to 91 GB/s
WRITE:	up to 65 GB/s
random 4K:	~ 1,x million IOPS

mmxcp (since 5.1.2)

Use the mmxcp command to perform parallel copies of files from a source directory to a target directory in a single IBM Spectrum Scale cluster. The copy can occur within a single file system or across different file systems in the same cluster. It can copy from a live file system or from a global or independent fileset snapshot. The mmxcp command has a strong relationship with the mmapplypolicy command.

Usage:

```
mmxcp enable --source Directory [--snapshot <Device:[FilesetName:]SnapshotName>] --target Directory
        [--force] [--copy-migrated] [-N {Node[,Node...] | NodeFile | NodeClass}]
        [-a IscanThreads] [-B MaxFiles] [-g GlobalWorkDirectory] [-L n] [-m ThreadLevel] [-n DirThreadLevel]
        [-s LocalWorkDirectory] [--sort-buffer-size Size] [--qos QOSClass]
or
mmxcp list {all | xcpID} [-Y]
or
mmxcp config {--get-max-value | --set-max-value Value}
```

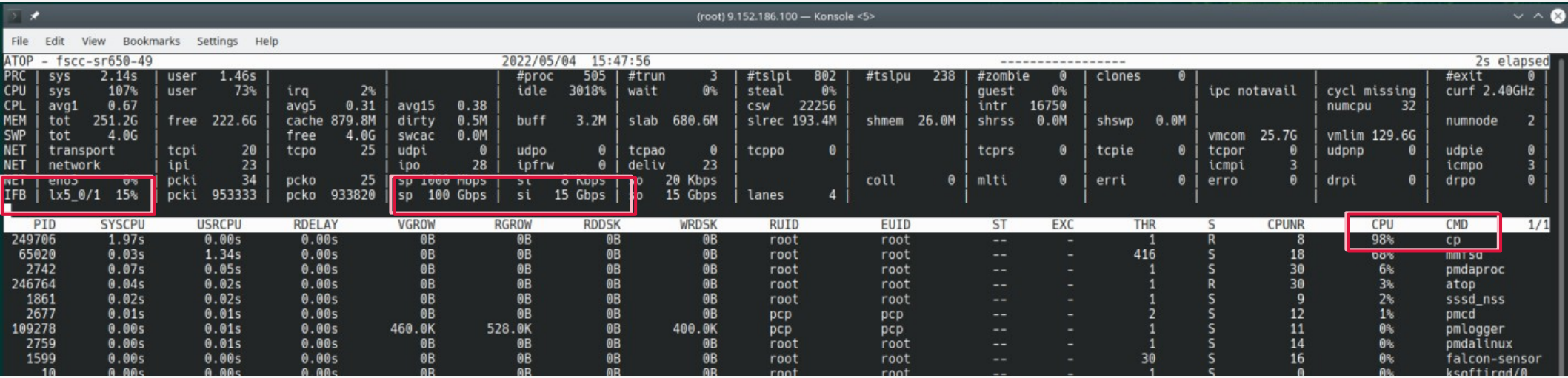
mmxcp- runs in parallel, using POSIX cp

```
[root@fscc-sr650-49 ~]#
time /usr/lpp/mmfs/bin/mmxcp enable --source /gpfs/beer8m/copy --target /gpfs/beer8m/copy2
[I] Beginning verification of parallel copy command parameters.
[I] Running policy commands with generated rules based on input configuration.
[I] Participating nodes will log information to their copy of /var/adm/ras/mmxcp.log.
[I] 2022-05-04@13:23:47.096 Directory entries scanned: 3.
[I] 2022-05-04@13:23:47.177 Parallel-piped sort and policy evaluation. 3 files scanned.
[I] 2022-05-04@13:23:47.190 Piped sorting and candidate file choosing. 3 records scanned.
[I] 2022-05-04@13:25:41.971 Policy execution. 3 files dispatched.
[I] Successfully completed running parallel copy command.
```

/usr/lpp/mmfs/bin/xcputil.sh LIST /gpfs/beer8m/.mmSharedTmpDir/mmPolicy.ix.203318.914DBA5A.1 %2Fgpfs%2Fbeer8m%
beer8m%2Fcopy2,false STARTING
2022-05-04_15:23:47.296+0200:203406:203426:/usr/lpp/mmfs/bin/xcputil.sh:185: Not copying file: /gpfs/beer8m/copy
[root@fscc-sr650-49 ~]# cat /gpfs/beer8m/.mmSharedTmpDir/mmPolicy.ix.203318.914DBA5A.1
265219 2136205783 0   -- /gpfs/beer8m/copy
5404745 1358718241 0   -- /gpfs/beer8m/copy/myfilefoo1
5404747 1181319670 0   -- /gpfs/beer8m/copy/myfilefoo2
[root@fscc-sr650-49 ~]#

```
366  0.0  0.0 152904 10348 ?        Ss   08:42   0:00  \_ sshd: root [priv]
368  0.0  0.0 152904  5428 ?        S    08:42   0:00  |   \_ sshd: root@pts/0
369  0.0  0.0  27060  5876 pts/0    Ss   08:42   0:00  |       \_ -bash
418  1.8  0.0  23404 12792 pts/0    S+   15:17   0:00  |           \_ /usr/lpp/mmfs/bin/mmksh /usr/lpp/mmfs/bin/mmxcp enable --source /gpfs/beer8m/copy --target /
000  2.2  0.0  19932 11812 pts/0    S+   15:17   0:00  |               \_ /usr/lpp/mmfs/bin/mmksh /usr/lpp/mmfs/bin/mmapplypolicy /gpfs/beer8m/copy -P /var/mmfs/t
121  0.0  0.0 2220048 7728 pts/0    Sl+  15:17   0:00  |                   \_ /usr/lpp/mmfs/bin/tsapolicy /gpfs/beer8m/copy -g /gpfs/beer8m/.mmSharedTmpDir -P /va
126  0.2  0.0 1943460 14760 pts/0   Sl+  15:17   0:00  |                       \_ /usr/lpp/mmfs/bin/tsapolicy /gpfs/beer8m/copy -g /gpfs/beer8m/.mmSharedTmpDir -P
226  0.0  0.0   9768  1152 pts/0    S+   15:17   0:00  |                           \_ sh -c /usr/lpp/mmfs/bin/xcputil.sh LIST '/gpfs/beer8m/.mmSharedTmpDir/mmPoli
242  1.2  0.0  19740 11592 pts/0    S+   15:17   0:00  |                               \_ /usr/lpp/mmfs/bin/mmksh /usr/lpp/mmfs/bin/xcputil.sh LIST /gpfs/beer8m/.
270 82.6  0.0  25872 10296 pts/0    R+   15:17   0:02  |                                   \_ /bin/cp -dR --parents --preserve=xattr,links,ownership,timestamps my
140  0.0  0.0 152904 10304 ?        Ss   13:46   0:00  \_ sshd: root [priv]
```

/bin/cp -dR --parents --preserve=xattr,links,ownership,timestamps myfilefoo2 /gpfs/beer8



The bandwidth for individual file copy is limited by 1 core (99% CPU)

```
real    3m53.444s
user    0m3.043s
sys     3m44.960s
```

1,7 GB/s for single large files

Note: with parallel sessions / directories, filesets, etc..., mmxcp scales to network line speed

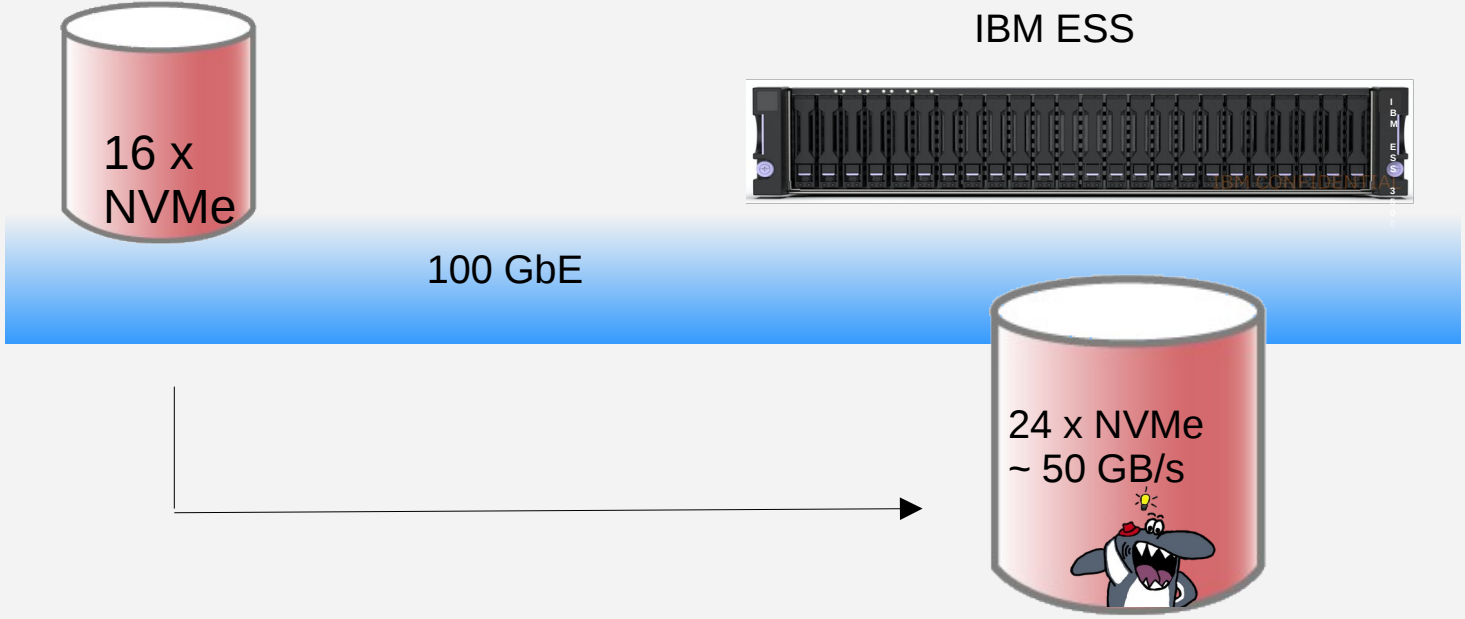
mmxcp (since 5.1.2)

– works great for copy data from GPFS ← → GPFS

– using POSIX cp tool from underlaying OS (Linux)

– high parallelism with multiple cp jobs managed by policy engine

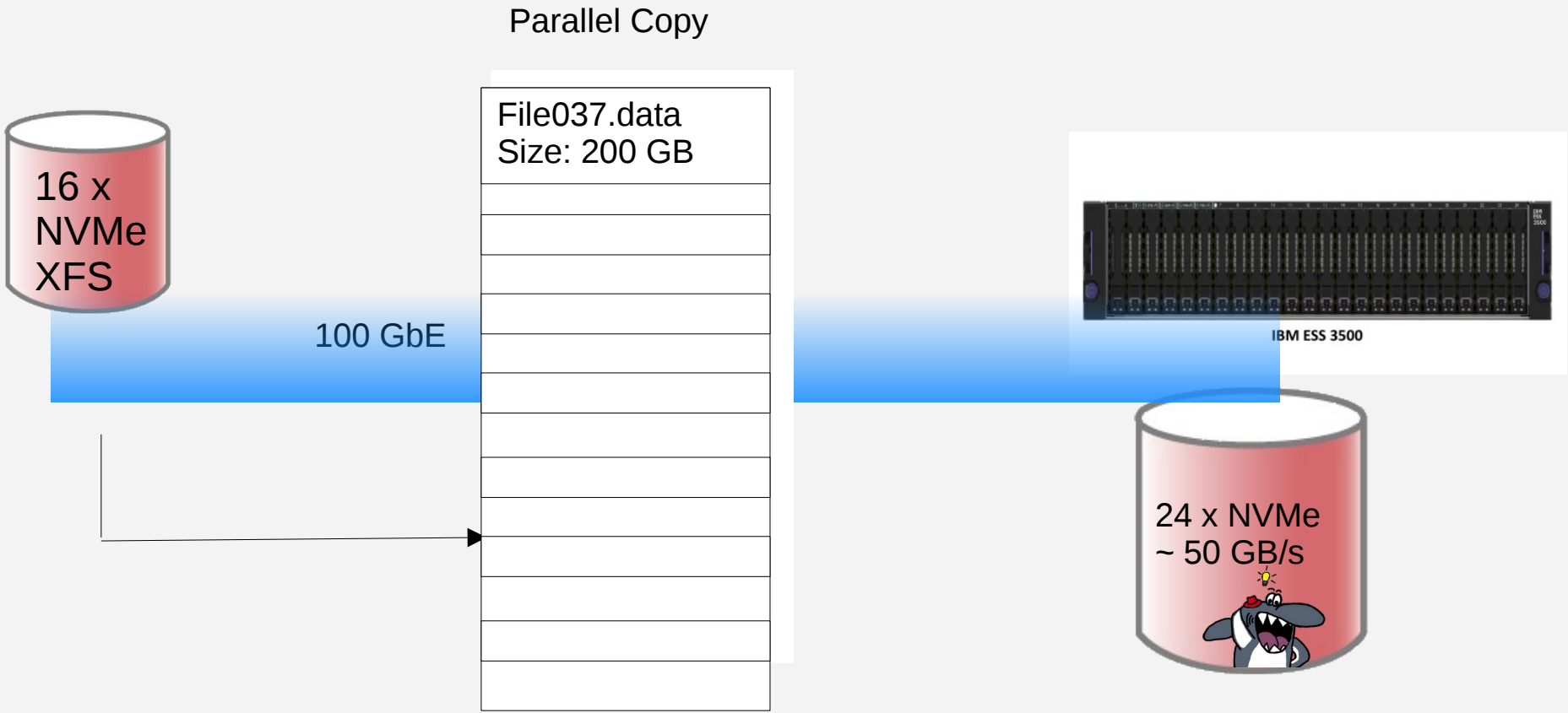
data ingest from existing data sources



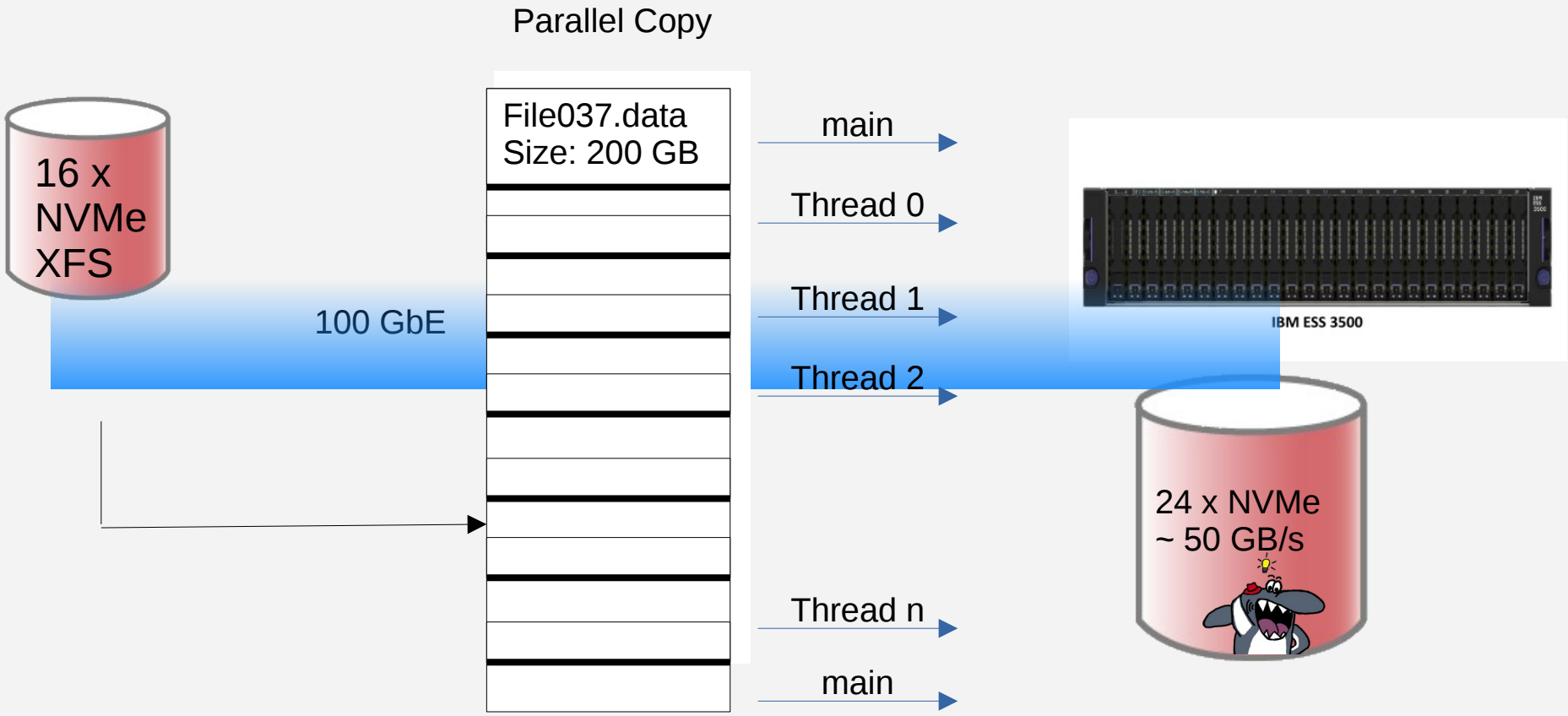
...a little research project…, parallel copy tool

Research project parallel copy

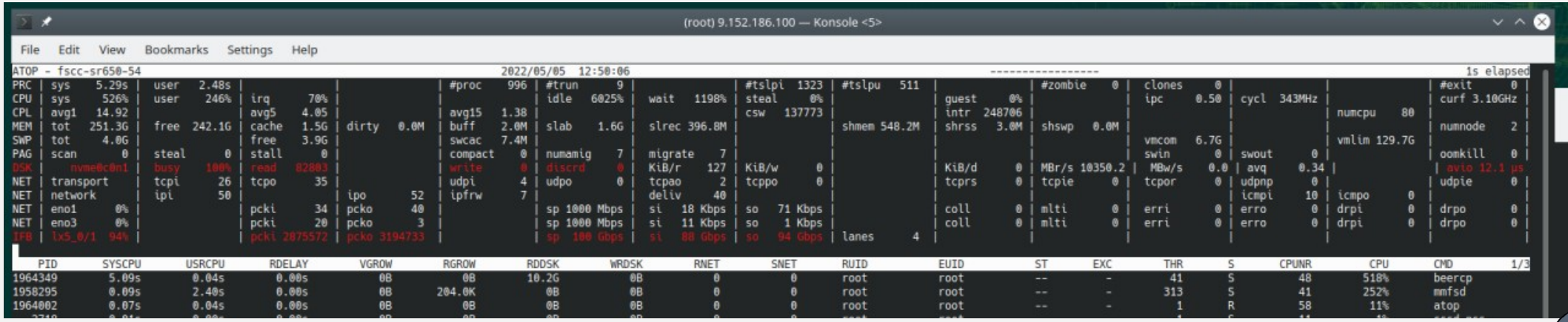
Research project parallel copy



Research project parallel copy

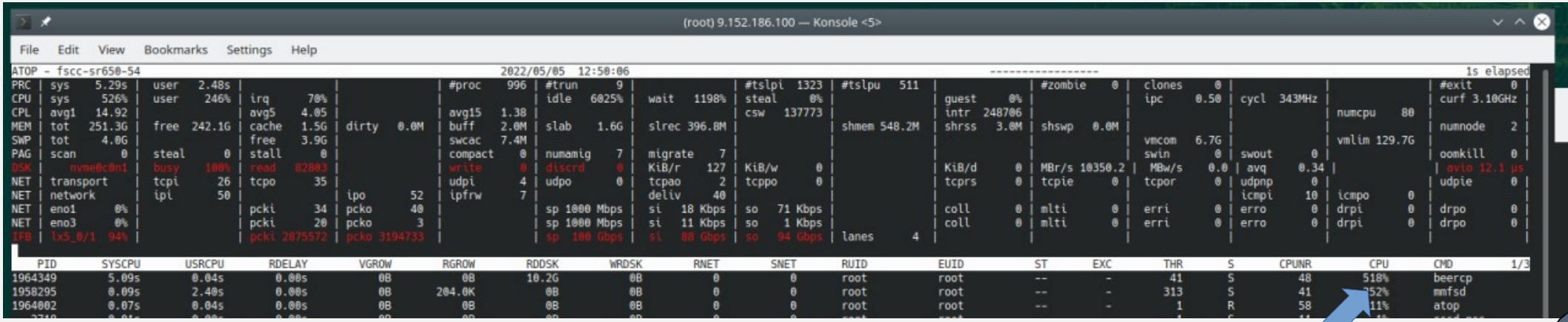


parallel (multi-threaded) copy

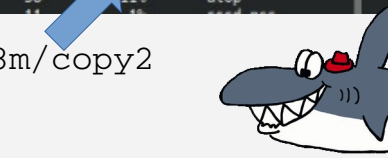



```
[root@fscc-sr650-54 mdlake]# time /root/beercp -p 40 -dio myfilefoo1 /gpfs/beer8m/copy2

real    0m22.559s
user    0m0.825s
sys     1m38.793s
```

parallel (multi-threaded) copy




```
[root@fscc-sr650-54 mdlake]# time /root/beercp -p 40 -dio myfilefoo1 /gpfs/beer8m/copy2

real    0m22.559s
user    0m0.825s
sys     1m38.793s

[root@fscc-sr650-54 mdlake]# ls -lh  myfilefoo1 /gpfs/beer8m/copy2/myfilefoo1
-rw-r--r-- 1 root root 200G May  5 12:50 /gpfs/beer8m/copy2/myfilefoo1
-rw-r--r-- 1 root root 200G Apr 27 23:50 myfilefoo1
[root@fscc-sr650-54 mdlake]#
[root@fscc-sr650-54 mdlake]# md5sum /gpfs/beer8m/copy2/myfilefoo1 ; md5sum myfilefoo1
4aceaaad68dccd5795063e35750f6b75  /gpfs/beer8m/copy2/myfilefoo1
4aceaaad68dccd5795063e35750f6b75  myfilefoo1
[root@fscc-sr650-54 mdlake]#
```



~ 9,x GB/s

Next - more details on the next UG / SSSD

- better prefetch
- Latency of small messages (ls -l .. stat())
- additional performance enhancements
  - driven by e.g. PCIe gen4 adjustments
- IO500 improvement