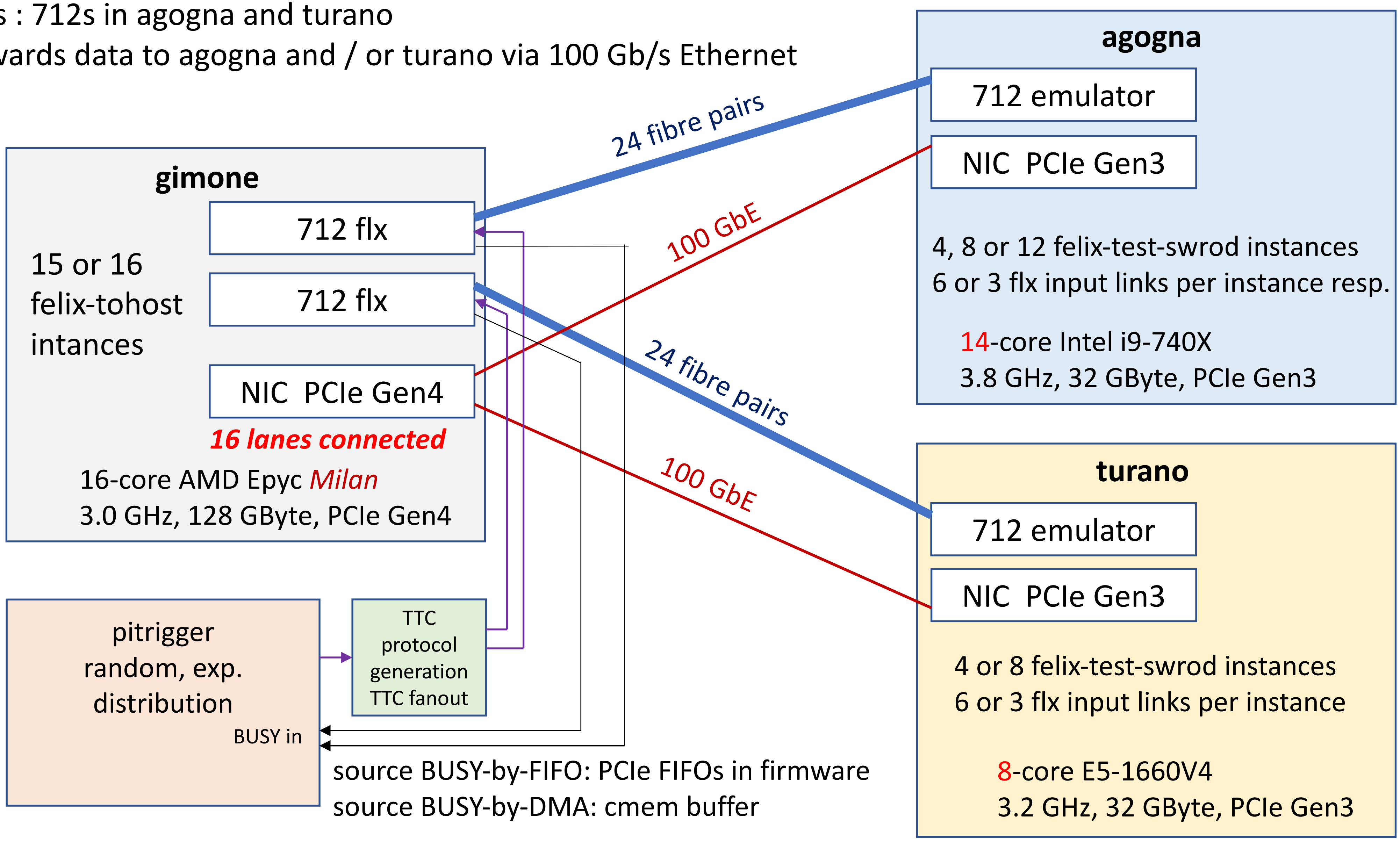
Test setup at Nikhef

FELIX server: gimone
data sources : 712s in agogna and turano
gimone forwards data to agogna and / or turano via 100 Gb/s Ethernet



gimone

712 flx

712 flx

15 or 16
felix-tohost
intances

NIC PCIe Gen4

16 lanes connected

16-core AMD Epyc *Milan*
3.0 GHz, 128 GByte, PCIe Gen4

24 fibre pairs

100 GbE

24 fibre pairs

100 GbE

agogna

712 emulator

NIC PCIe Gen3

4, 8 or 12 felix-test-swrod instances
6 or 3 flx input links per instance resp.

14-core Intel i9-740X
3.8 GHz, 32 GByte, PCIe Gen3

turano

712 emulator

NIC PCIe Gen3

4 or 8 felix-test-swrod instances
6 or 3 flx input links per instance

8-core E5-1660V4
3.2 GHz, 32 GByte, PCIe Gen3

pitrigger
random, exp.
distribution

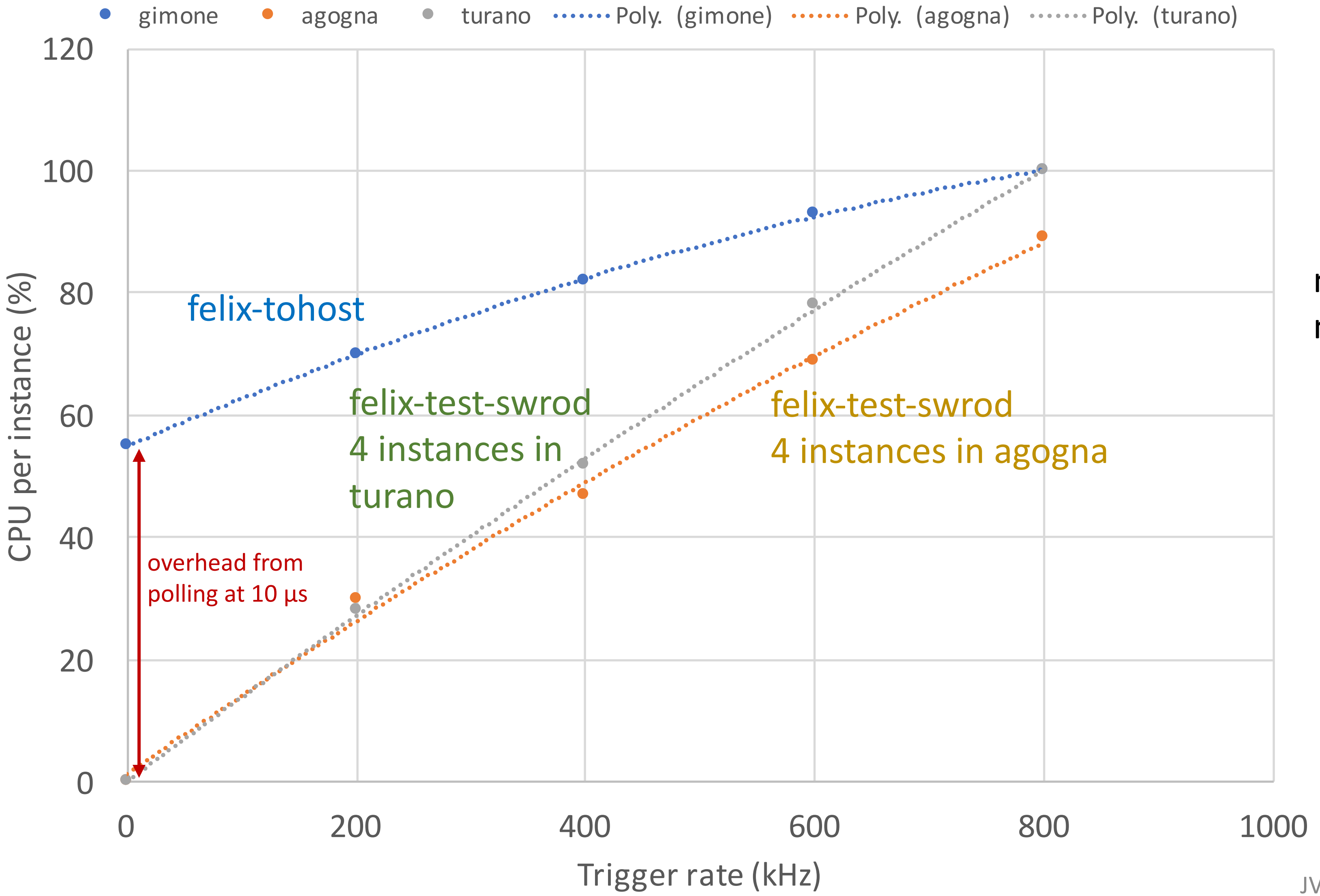
BUSY in

TTC
protocol
generation
TTC fanout

source BUSY-by-FIFO: PCIe FIFOs in firmware
source BUSY-by-DMA: cmem buffer

GBT mode, BUSY-by-FIFO off (no truncations), BUSY-by-DMA on, 0% dead time for rate ≤ 800 kHz for 15 felix-tohost instances

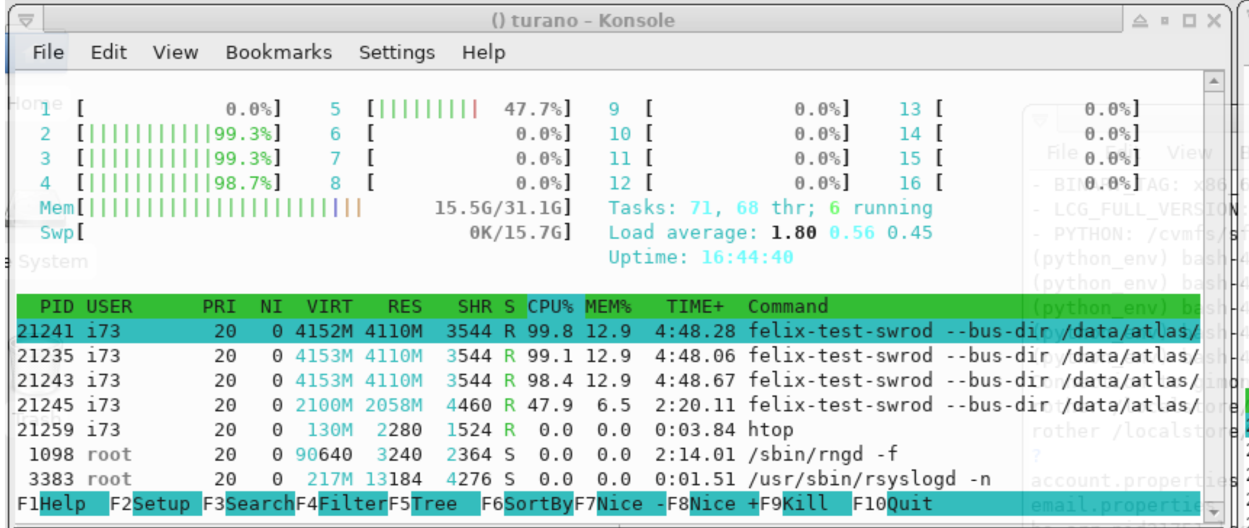
- eight 8-bit E-links per GBT link, 48 GBT links, 3 GBT links per DMA channel, 16 DMA channels in total, 20 bytes per trigger per E-link
- 16 felix-tohost instances, one per core: instable running, instance pinned to core 0 cannot handle the data
- 15 felix-tohost instances, one per core, servicing 45 GBT links: stable running if 4 felix-test-swrod instances on turano are pinned to core 1, 2, 3 and 4.
- Max. rate about 800 kHz: felix-tohost ~100% CPU, felix-test-swrod turano ~100% CPU, felix-test-swrod agogna ~88-90% CPU



microcode.service masked
mitigations = off

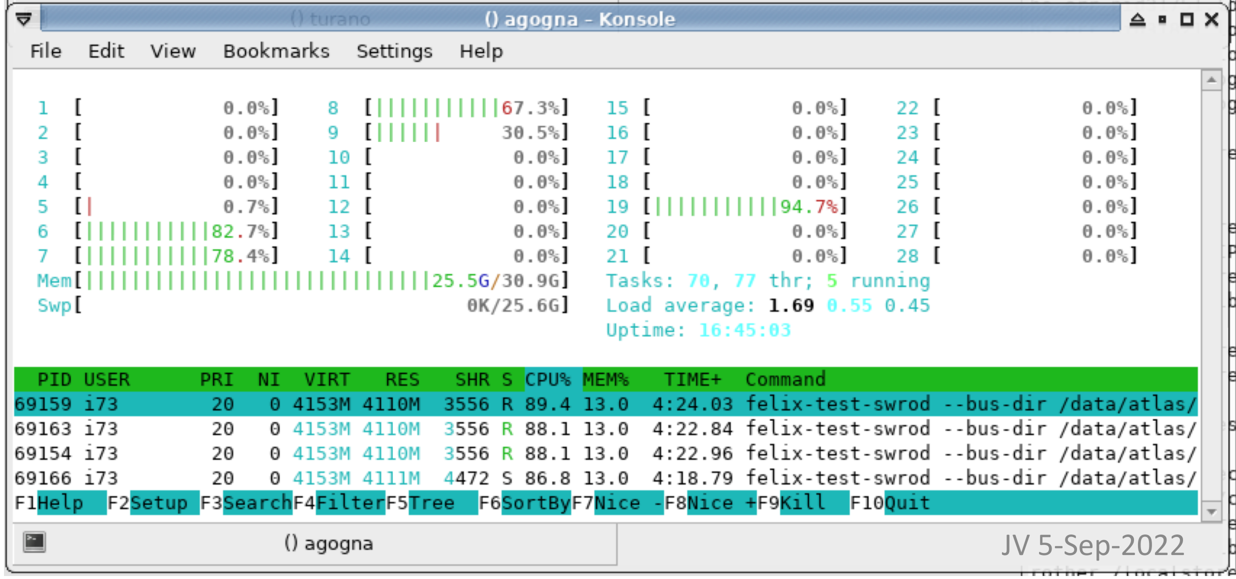
GBT mode, BUSY-by-FIFO off (no truncations), BUSY-by-DMA on, 0% dead time for rate ≤ 800 kHz for 15 felix-tohost instances

- eight 8-bit E-links per GBT link, 48 GBT links, 3 GBT links per DMA channel, 16 DMA channels in total, 20 bytes per trigger per E-link
- 16 felix-tohost instances, one per core: instable running, instance pinned to core 0 cannot handle the data
- 15 felix-tohost instances, one per core, servicing 45 GBT links: stable running if 4 felix-test-swrod instances on turano are pinned to core 1, 2, 3 and 4.
- Max. rate about 800 kHz: felix-tohost ~100% CPU, felix-test-swrod turano ~100% CPU, felix-test-swrod agogna ~88-90% CPU



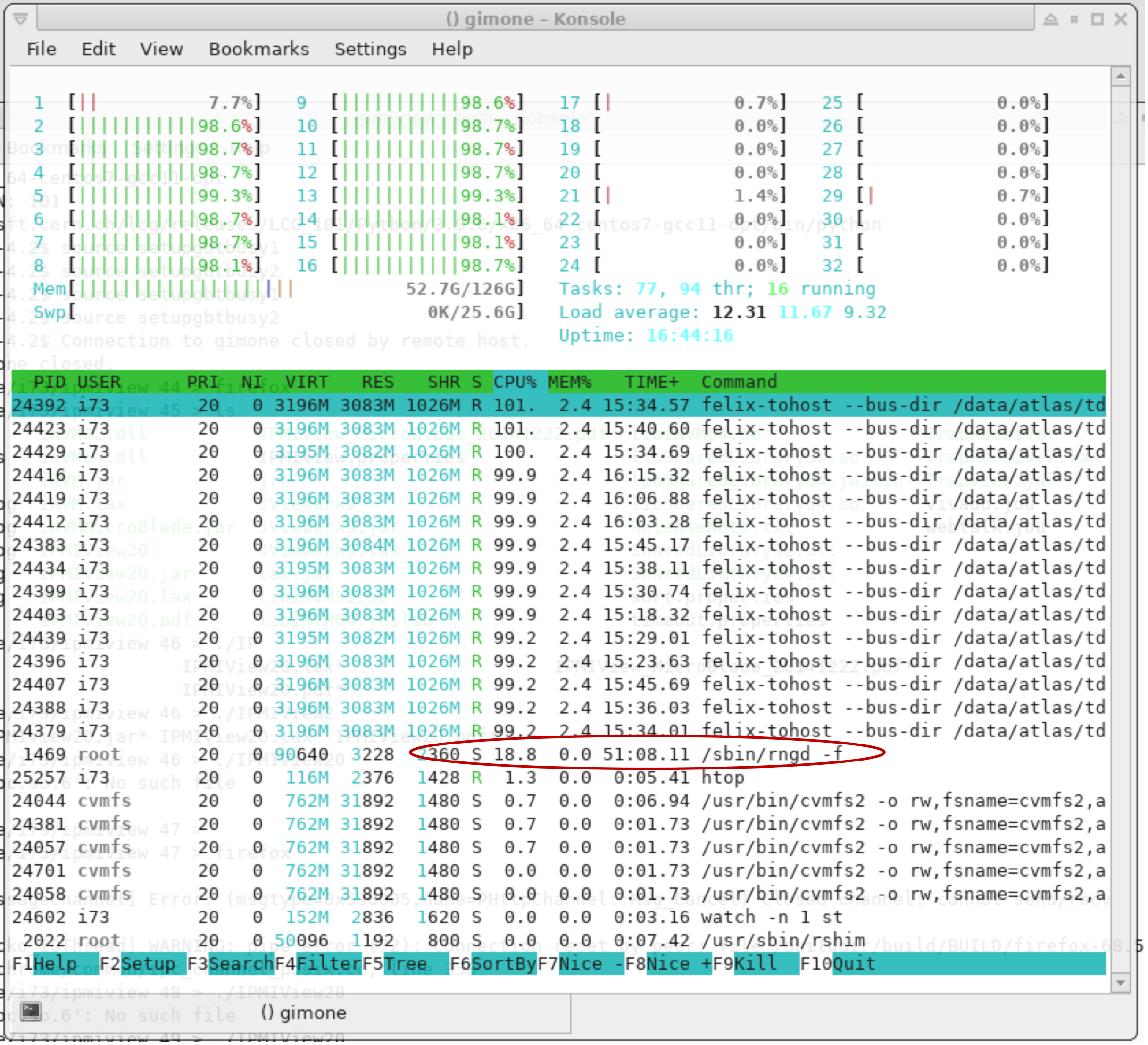
```
  1  [                    0.0%]    5  [|||||||||   47.7%]    9  [                   0.0%]    13 [                   0.0%]
  2  [|||||||||||99.3%]            6  [             0.0%]    10 [                   0.0%]    14 [                   0.0%]
  3  [|||||||||||99.3%]            7  [             0.0%]    11 [                   0.0%]    15 [                   0.0%]
  4  [|||||||||||98.7%]            8  [             0.0%]    12 [                   0.0%]    16 [                   0.0%]
  Mem[||||||||||||||||||||||||||   15.5G/31.1G]    Tasks: 71, 68 thr; 6 running
  Swp[                                0K/15.7G]    Load average: 1.80 0.56 0.45
                                                   Uptime: 16:44:40

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
21241 i73        20   0 4152M 4110M  3544 R 99.8 12.9  4:48.28 felix-test-swrod --bus-dir /data/atlas/
21235 i73        20   0 4153M 4110M  3544 R 99.1 12.9  4:48.06 felix-test-swrod --bus-dir /data/atlas/
21243 i73        20   0 4153M 4110M  3544 R 98.4 12.9  4:48.67 felix-test-swrod --bus-dir /data/atlas/
21245 i73        20   0 2100M 2058M  4460 R 47.9  6.5  2:20.11 felix-test-swrod --bus-dir /data/atlas/
21259 i73        20   0  130M  2280  1524 R  0.0  0.0  0:03.84 htop
 1098 root       20   0 90640  3240  2364 S  0.0  0.0  2:14.01 /sbin/rngd -f
 3383 root       20   0  217M 13184  4276 S  0.0  0.0  0:01.51 /usr/sbin/rsyslogd -n
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```



```
  1  [                 0.0%]    8  [|||||||||||67.3%]    15 [                 0.0%]    22 [                 0.0%]
  2  [                 0.0%]    9  [||||||      30.5%]    16 [                 0.0%]    23 [                 0.0%]
  3  [                 0.0%]    10 [             0.0%]    17 [                 0.0%]    24 [                 0.0%]
  4  [                 0.0%]    11 [             0.0%]    18 [                 0.0%]    25 [                 0.0%]
  5  [|                0.7%]    12 [             0.0%]    19 [|||||||||||94.7%]      26 [                 0.0%]
  6  [|||||||||||82.7%]         13 [             0.0%]    20 [                 0.0%]    27 [                 0.0%]
  7  [|||||||||||78.4%]         14 [             0.0%]    21 [                 0.0%]    28 [                 0.0%]
  Mem[|||||||||||||||||||||||||||||||25.5G/30.9G]    Tasks: 70, 77 thr; 5 running
  Swp[                              0K/25.6G]    Load average: 1.69 0.55 0.45
                                                 Uptime: 16:45:03

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
69159 i73        20   0 4153M 4110M  3556 R 89.4 13.0  4:24.03 felix-test-swrod --bus-dir /data/atlas/
69163 i73        20   0 4153M 4110M  3556 R 88.1 13.0  4:22.84 felix-test-swrod --bus-dir /data/atlas/
69154 i73        20   0 4153M 4110M  3556 R 88.1 13.0  4:22.96 felix-test-swrod --bus-dir /data/atlas/
69166 i73        20   0 4153M 4111M  4472 S 86.8 13.0  4:18.79 felix-test-swrod --bus-dir /data/atlas/
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```



```
  1  [||          7.7%]    9  [|||||||||||98.6%]    17 [|                0.7%]    25 [                 0.0%]
  2  [|||||||||||98.6%]    10 [|||||||||||98.7%]    18 [                 0.0%]    26 [                 0.0%]
  3  [|||||||||||98.7%]    11 [|||||||||||98.7%]    19 [                 0.0%]    27 [                 0.0%]
  4  [|||||||||||98.7%]    12 [|||||||||||98.7%]    20 [                 0.0%]    28 [                 0.0%]
  5  [|||||||||||99.3%]    13 [|||||||||||99.3%]    21 [|                1.4%]    29 [|                0.7%]
  6  [|||||||||||98.7%]    14 [|||||||||||98.1%]    22 [                 0.0%]    30 [                 0.0%]
  7  [|||||||||||98.7%]    15 [|||||||||||98.1%]    23 [                 0.0%]    31 [                 0.0%]
  8  [|||||||||||98.1%]    16 [|||||||||||98.7%]    24 [                 0.0%]    32 [                 0.0%]
  Mem[|||||||||||||||||||||    52.7G/126G]    Tasks: 77, 94 thr; 16 running
  Swp[                          0K/25.6G]    Load average: 12.31 11.67 9.32
                                             Uptime: 16:44:16

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
24392 i73        20   0 3196M 3083M 1026M R 101.  2.4 15:34.57 felix-tohost --bus-dir /data/atlas/td
24423 i73        20   0 3196M 3083M 1026M R 101.  2.4 15:40.60 felix-tohost --bus-dir /data/atlas/td
24429 i73        20   0 3195M 3082M 1026M R 100.  2.4 15:34.69 felix-tohost --bus-dir /data/atlas/td
24416 i73        20   0 3196M 3083M 1026M R 99.9  2.4 16:15.32 felix-tohost --bus-dir /data/atlas/td
24419 i73        20   0 3196M 3083M 1026M R 99.9  2.4 16:06.88 felix-tohost --bus-dir /data/atlas/td
24412 i73        20   0 3196M 3083M 1026M R 99.9  2.4 16:03.28 felix-tohost --bus-dir /data/atlas/td
24383 i73        20   0 3196M 3084M 1026M R 99.9  2.4 15:45.17 felix-tohost --bus-dir /data/atlas/td
24434 i73        20   0 3195M 3083M 1026M R 99.9  2.4 15:38.11 felix-tohost --bus-dir /data/atlas/td
24399 i73        20   0 3196M 3083M 1026M R 99.9  2.4 15:30.74 felix-tohost --bus-dir /data/atlas/td
24403 i73        20   0 3196M 3083M 1026M R 99.9  2.4 15:18.32 felix-tohost --bus-dir /data/atlas/td
24439 i73        20   0 3195M 3082M 1026M R 99.2  2.4 15:29.01 felix-tohost --bus-dir /data/atlas/td
24396 i73        20   0 3196M 3083M 1026M R 99.2  2.4 15:23.63 felix-tohost --bus-dir /data/atlas/td
24407 i73        20   0 3196M 3083M 1026M R 99.2  2.4 15:45.69 felix-tohost --bus-dir /data/atlas/td
24388 i73        20   0 3196M 3083M 1026M R 99.2  2.4 15:36.03 felix-tohost --bus-dir /data/atlas/td
24379 i73        20   0 3196M 3083M 1026M R 99.2  2.4 15:34.01 felix-tohost --bus-dir /data/atlas/td
 1469 root       20   0 90640  3228  2360 S 18.8  0.0 51:08.11 /sbin/rngd -f
25257 i73        20   0  116M  2376  1428 R  1.3  0.0  0:05.41 htop
24044 cvmfs      20   0  762M 31892  1480 S  0.7  0.0  0:06.94 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,a
24381 cvmfs      20   0  762M 31892  1480 S  0.7  0.0  0:01.73 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,a
24057 cvmfs      20   0  762M 31892  1480 S  0.7  0.0  0:01.73 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,a
24701 cvmfs      20   0  762M 31892  1480 S  0.0  0.0  0:01.73 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,a
24058 cvmfs      20   0  762M 31892  1480 S  0.0  0.0  0:01.73 /usr/bin/cvmfs2 -o rw,fsname=cvmfs2,a
24602 i73        20   0  152M  2836  1620 S  0.0  0.0  0:03.16 watch -n 1 st
 2022 root       20   0 50096  1192   800 S  0.0  0.0  0:07.42 /usr/sbin/rshim
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

JV 5-Sep-2022

Maximum rate GBT mode determined by:

- felix-to-host CPU usage per instance per instance 100% at 800 kHz, but polling interval can be made larger:
  - Polling interval 10 µs: 55% CPU usage for 0 kHz trigger rate
  - Polling interval 100 µs: 5.8 – 7.1% CPU usage for 0 kHz trigger rate
  - Polling interval 1 ms: 0.7% CPU usage for 0 kHz trigger rate

- 100% CPU not possible for 16th felix-to-host instance, but increase of polling interval length also reduces Linux overhead:
  - Polling interval 10 µs: /sbin/rngd –f: 18.8%
  - Polling interval 100 µs: /sbin/rngd –f: 4.2%
  - Polling interval 1 ms: /sbin/rngd –f: 0.7%

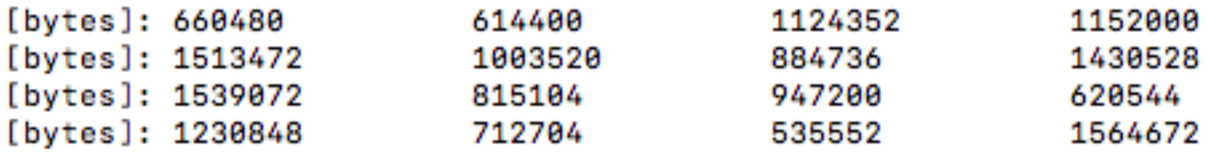
- felix-test-swrod runs on turano at 100% CPU at 800 kHz, handling data from 6 GBT links / 48 E-links
  - ➢ 16 instances and therefore 16 cores needed if 3 GBT links per felix-test-swrod instance are handled for reduction of CPU usage
  - ➢ turano has 8 cores and -> run at max. 7 instances, maybe 8 (if allowed by Linux overhead)
  - ➢ Tests done with 12 instances on agogna (14-core CPU) and 4 on turano
  - ➢ Could have tested with less instances on agogna and more on turano, would have resulted in more even bandwidth usage of both 100 GbE links (short test of this afternoon: 8 instances on agogna and 8 on turano: same result as for 12 instances on agogna and 4 on turano)

Result with 16 felix-to-host instances, 28 bytes per E-link per trigger for first 712, 24 bytes per E-link for second 712:
1 MHz rate just possible with 0% BUSY-by-DMA during 1 hour. 28 bytes per E-link for second 712 causes sometimes BUSY-by-DMA. felix-to-host for last link DMA channel of second 712 runs on core 0.
Bandwidth usage reported by felix-test-swrod : 55 Gb/s for 100 GbE link to agogna, 16 Gb/s for 100 GbE link to turano

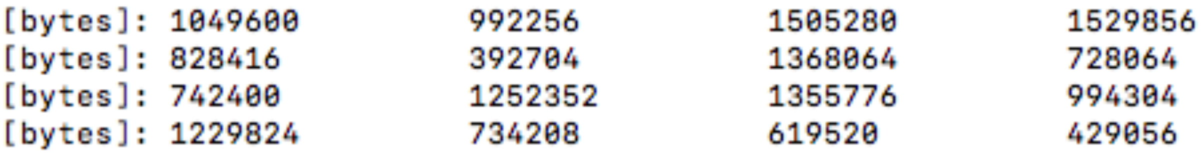
FULL mode tests done in buffered mode, polling interval of 100 us

- Mellanox software used, previous observed low maximum rates not seen again
- BUSY-by-FIFO switched off, no truncations
- 4 felix-test-swrod instances on agogna and 4 on turano, each handling data from 6 flx input links
- felix-tohost instances running below ~ 45% CPU
- felix-test-swrod instances running at about 30% CPU
- Max. rate ~ 1050 kHz, 0% BUSY-by-DMA
- about 72 Gb/s via each 100 GbE link (PCIe Gen 4 interface!): network bandwidth seems to limit the rate, CPU usage felix-tohost not if polling interval is not too short (~100% at 10 us felix-tohost)

Two snapshots of cmem buffer (1 GB per felix-tohost instance) usage in bytes @ 1050 KHz

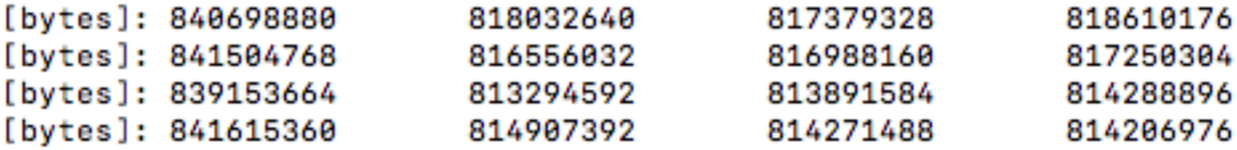


```
device 0  [bytes]: 660480     614400     1124352    1152000
device 1  [bytes]: 1513472    1003520    884736     1430528
device 2  [bytes]: 1539072    815104     947200     620544
device 3  [bytes]: 1230848    712704     535552     1564672
```



```
device 0  [bytes]: 1049600    992256     1505280    1529856
device 1  [bytes]: 828416     392704     1368064    728064
device 2  [bytes]: 742400     1252352    1355776    994304
device 3  [bytes]: 1229824    734208     619520     429056
```

Snapshot of cmem buffer (1 GB per felix-tohost instance) usage in bytes @ 1100 KHz after 30 s



```
device 0  [bytes]: 840698880    818032640    817379328    818610176
device 1  [bytes]: 841504768    816556032    816988160    817250304
device 2  [bytes]: 839153664    813294592    813891584    814288896
device 3  [bytes]: 841615360    814907392    814271488    814206976
```

-> BUSY-by-DMA (threshold @ about 800 MB)

Remarks:

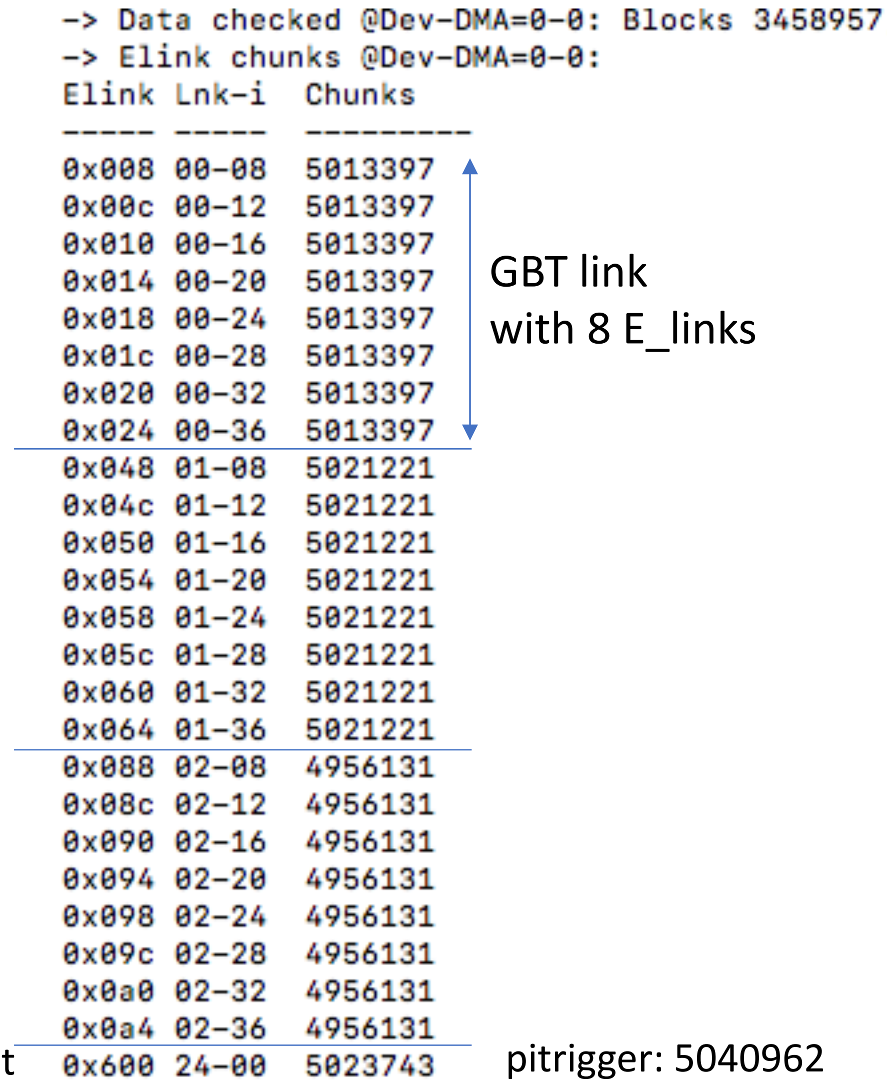
- Earlier report: small dead time fraction at high rates seems to be difficult to avoid.
    - Dead time fraction observed caused by BUSY-by-FIFO, thresholds for setting and removing the BUSY were set, upon the basis of trial and error, to avoid truncations
    - Thresholds used effective for avoiding truncations, but cause a small dead-time fraction also at 1 MHz or lower rates
    - BUSY-by-FIFO has been disabled as no truncation occurs with the firmware currently used and if the firmware is properly functioning.
    - Resetting the firmware or rebooting the server is sometimes necessary, in particular for the GBT mode firmware, to get the firmware properly functioning

- The FELIG emulator consistently outputs a somewhat smaller number of event fragments than the number of triggers, as seen in the totals reported by fdaqm. For different GBT links the number of fragments may be somewhat different (of the order of 0.05%), but for the E-links of one GBT link they are the same, example on next slide

- Pinning of felix-tohost and felix-test-swrod instances to cores (with taskset) ineeded for best performance if CPU load of each instance approaches 100%. If the CPU load is less pinning may not be needed, in that case reduction of /proc/sys/kernel/sched_migration_cost_ns, as found for the ROS, could help. Some experimentation (GBT mode, 1 ms polling, 1 MHz rate) with setting it to 1000 instead of 50000 without core pinning to felix-tohost instances did not show a clear improvement, and peaks in cmem buffer filling appeared to be considerably higher than for pinning, caused sometimes BUSY-by-DMA

(From RHEL7 doc: /proc/sys/kernel/sched_migration_cost_ns specifies the amount of time after the last execution that a task is considered to be “cache hot” in migration decisions. Increasing this variable reduces task migrations. Adjust by factor of 2-10x. Task migrations may be irrelevant depending on any configured task affinity settings).

5 s



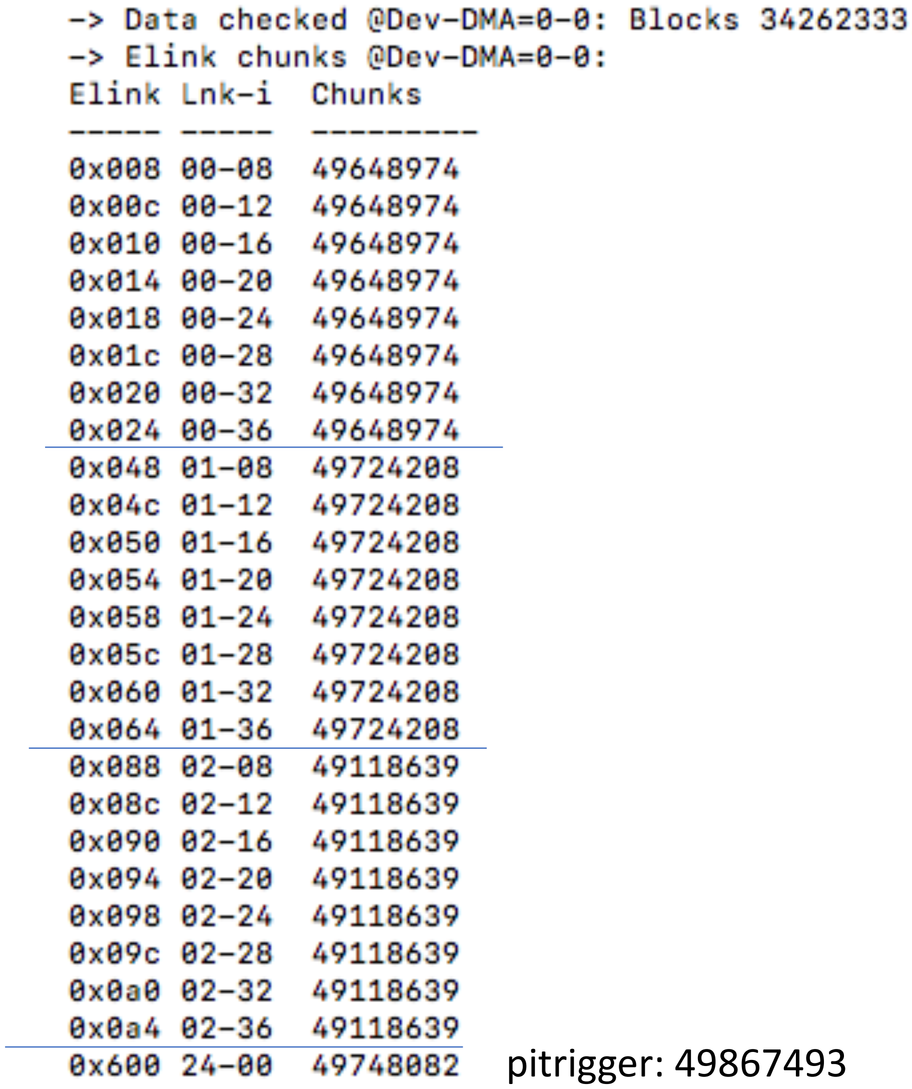
```
-> Data checked @Dev-DMA=0-0: Blocks 3458957
-> Elink chunks @Dev-DMA=0-0:
Elink Lnk-i  Chunks
----- -----  ---------
0x008 00-08  5013397
0x00c 00-12  5013397
0x010 00-16  5013397
0x014 00-20  5013397
0x018 00-24  5013397
0x01c 00-28  5013397
0x020 00-32  5013397
0x024 00-36  5013397
0x048 01-08  5021221
0x04c 01-12  5021221
0x050 01-16  5021221
0x054 01-20  5021221
0x058 01-24  5021221
0x05c 01-28  5021221
0x060 01-32  5021221
0x064 01-36  5021221
0x088 02-08  4956131
0x08c 02-12  4956131
0x090 02-16  4956131
0x094 02-20  4956131
0x098 02-24  4956131
0x09c 02-28  4956131
0x0a0 02-32  4956131
0x0a4 02-36  4956131
0x600 24-00  5023743
```

GBT link
with 8 E_links

TTC-to-host

pitrigger: 5040962

50 s



```
-> Data checked @Dev-DMA=0-0: Blocks 34262333
-> Elink chunks @Dev-DMA=0-0:
Elink Lnk-i  Chunks
----- -----  ---------
0x008 00-08  49648974
0x00c 00-12  49648974
0x010 00-16  49648974
0x014 00-20  49648974
0x018 00-24  49648974
0x01c 00-28  49648974
0x020 00-32  49648974
0x024 00-36  49648974
0x048 01-08  49724208
0x04c 01-12  49724208
0x050 01-16  49724208
0x054 01-20  49724208
0x058 01-24  49724208
0x05c 01-28  49724208
0x060 01-32  49724208
0x064 01-36  49724208
0x088 02-08  49118639
0x08c 02-12  49118639
0x090 02-16  49118639
0x094 02-20  49118639
0x098 02-24  49118639
0x09c 02-28  49118639
0x0a0 02-32  49118639
0x0a4 02-36  49118639
0x600 24-00  49748082
```

pitrigger: 49867493