

# Square Pegs in Round holes

or System Level Performance Data and perf

Paweł Moll <pawel.moll@arm.com>

# The plan

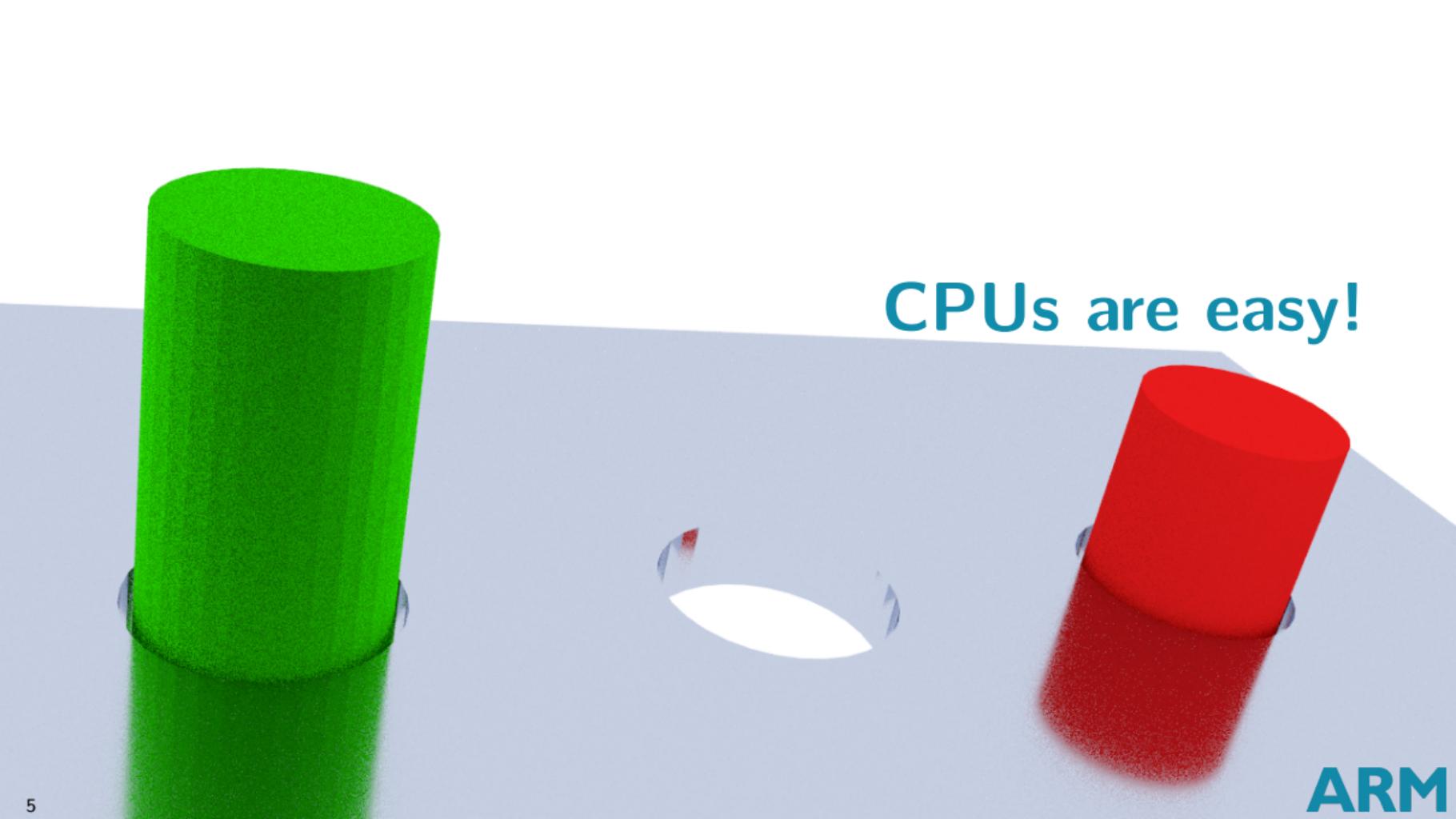
- Problem definition
- CPUs
- Systems
- perf and non-CPUs
- Examples

# Disclaimer

- perf is growing...
  - ... it's huge now...
  - ... but its documentation is... poor ;-)
- I'm far far far from being an expert
- I'm declaring a bounty for "it's already solved!" comment

# A program is running slow

- need faster CPU?
- profile, find hotspots
- better algorithm
- micro-architectural optimisations

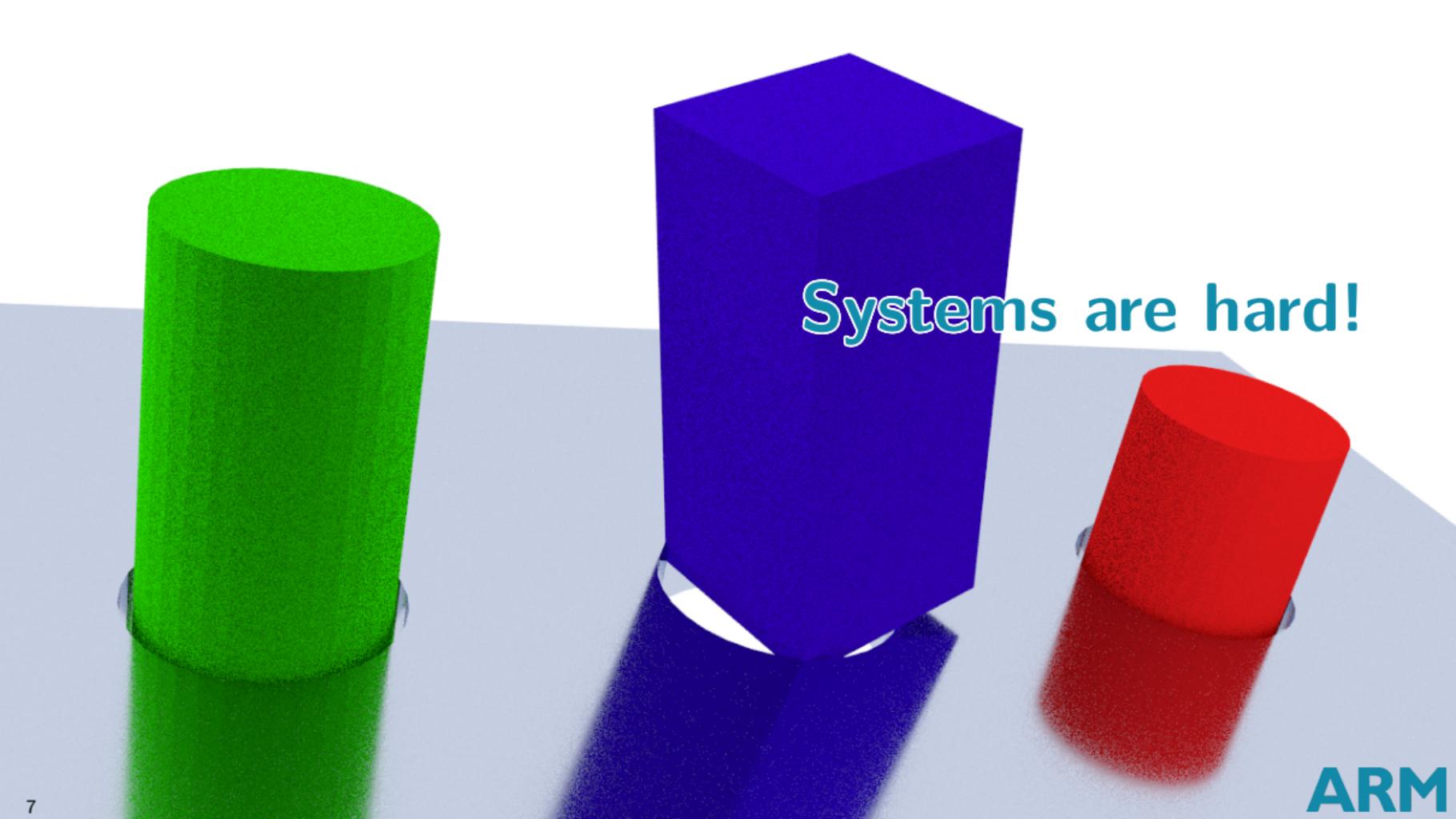


CPUs are easy!

ARM

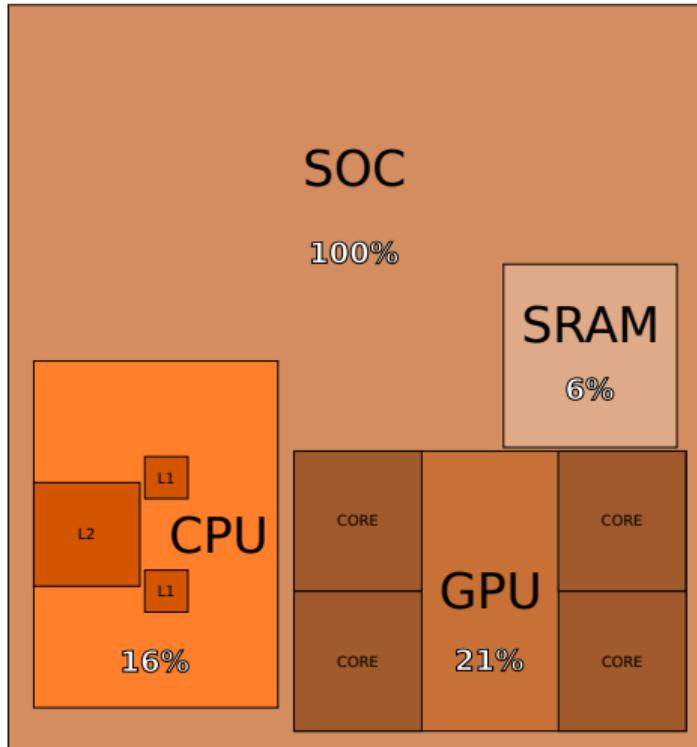
# CPU (software) performance

- Textbook knowledge
  - Taught from year one
- Well understood
  - Even cache effects
- Widely available tools

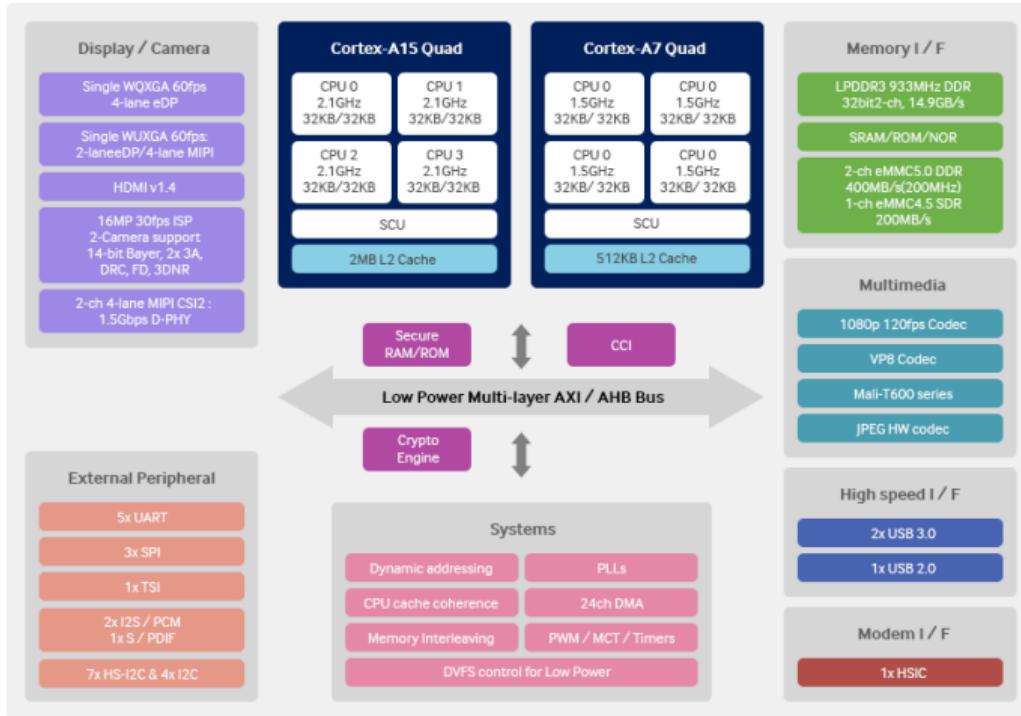


Systems are hard!

# CPU is a tiny piece of the puzzle



# CPU is just one of the players



source: [www.samsung.com](http://www.samsung.com)

# Bandwidth

- Limited resource
- Many consumers

## Example issue:

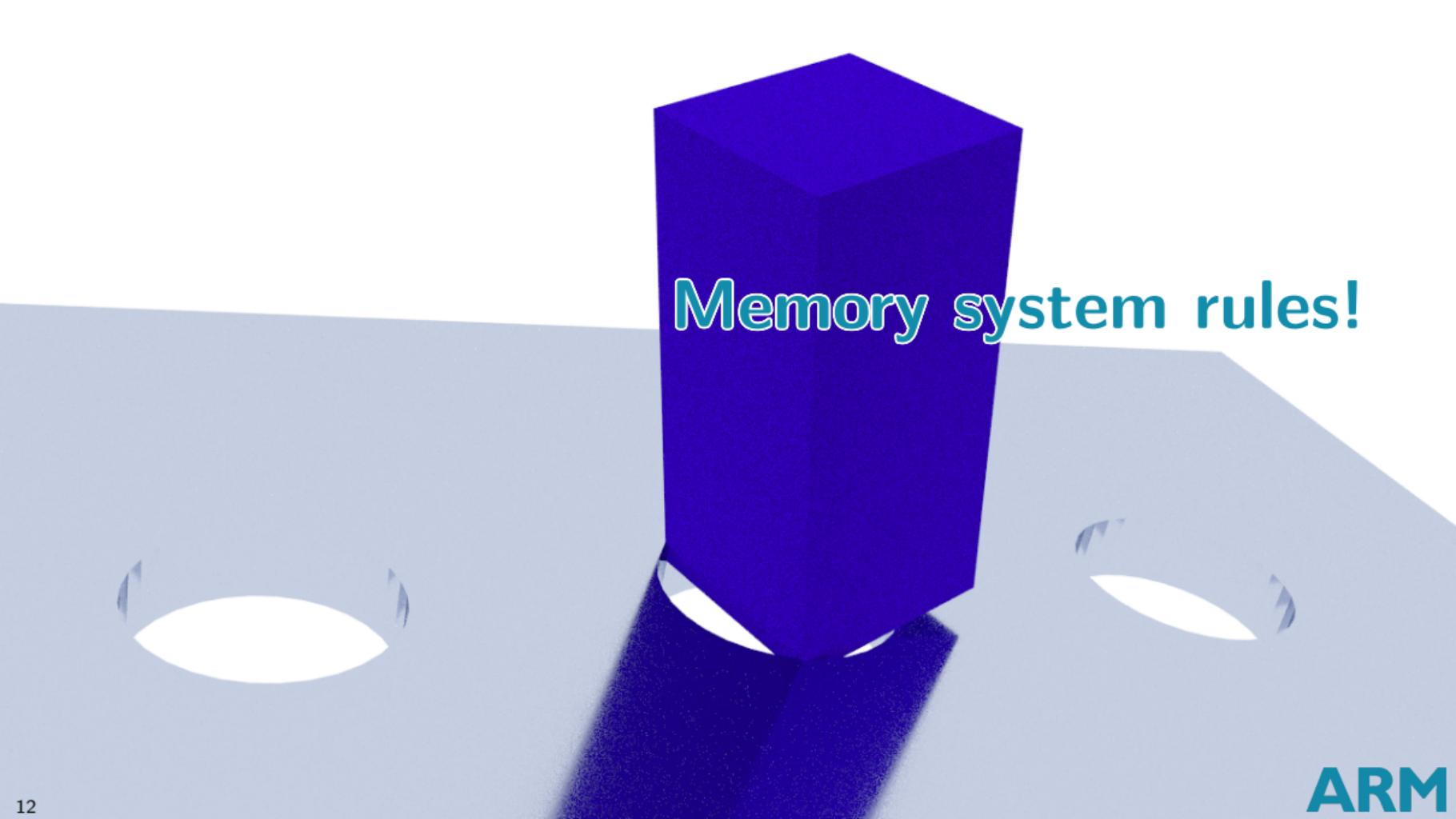
- Processing 300 byte network packet
- 5MB of memory traffic (yes, 20000x!)

# Latency

- CPU cycle at 2GHz:  $0.5\text{ns}$
- memory access latency:  $50 - 150\text{ns}$

## Example issue:

- Asynchronous bridges
- Up to 20% of latency



**Memory system rules!**

# Performance analysis tools

- perf
- valgrind (cachegrind)
- gem5
- spreadsheet

# CONFIG\_PERF\_EVENTS

- `man perf_event_open`
- CPU centric
- Popular subject at LinuxCon:
  - Perf & CTF - Jiri Olsa
  - Performance Analysis Using the Perf Suite - Mans Rullgard
  - Linux Performance Tools - Brendan Gregg
  - Cycle Accurate Profiling With Perf - Paweł Moll

## CONFIG\_PERF\_EVENTS

```
$ perf stat - make -j
```

Performance counter stats for 'make -j':

8117.370256	task clock ticks	#	11.281	CPU utilization factor
678	context switches	#	0.000	M/sec
133	CPU migrations	#	0.000	M/sec
235724	pagefaults	#	0.029	M/sec
24821162526	CPU cycles	#	3057.784	M/sec
18687303457	instructions	#	2302.138	M/sec
172158895	cache references	#	21.209	M/sec
27075259	cache misses	#	3.335	M/sec

Wall-clock time elapsed: 719.554352 msecs

## x86 uncore PMU

- Collection of package-level performance counters

```
# perf stat -a -C 0 -e 'uncore_cbox_0/clockticks/' sleep 1
```

Performance counter stats for 'system wide':

```
223,256,554 uncore_cbox_0/clockticks/
```

```
1.001035815 seconds time elapsed
```

### Possible improvement:

- Documentation! :-)

## x86 uncore PMU

- Interpretation (<http://lwn.net/Articles/536034/>)

```
# perf uncore -T
#-----
#           |          Socket0          |
#           |-----|
#   Time   |          RAM Bandwidth    |
#   in     |              Wr            Rd|
#   secs   |              MB/s          MB/s|
#-----
```

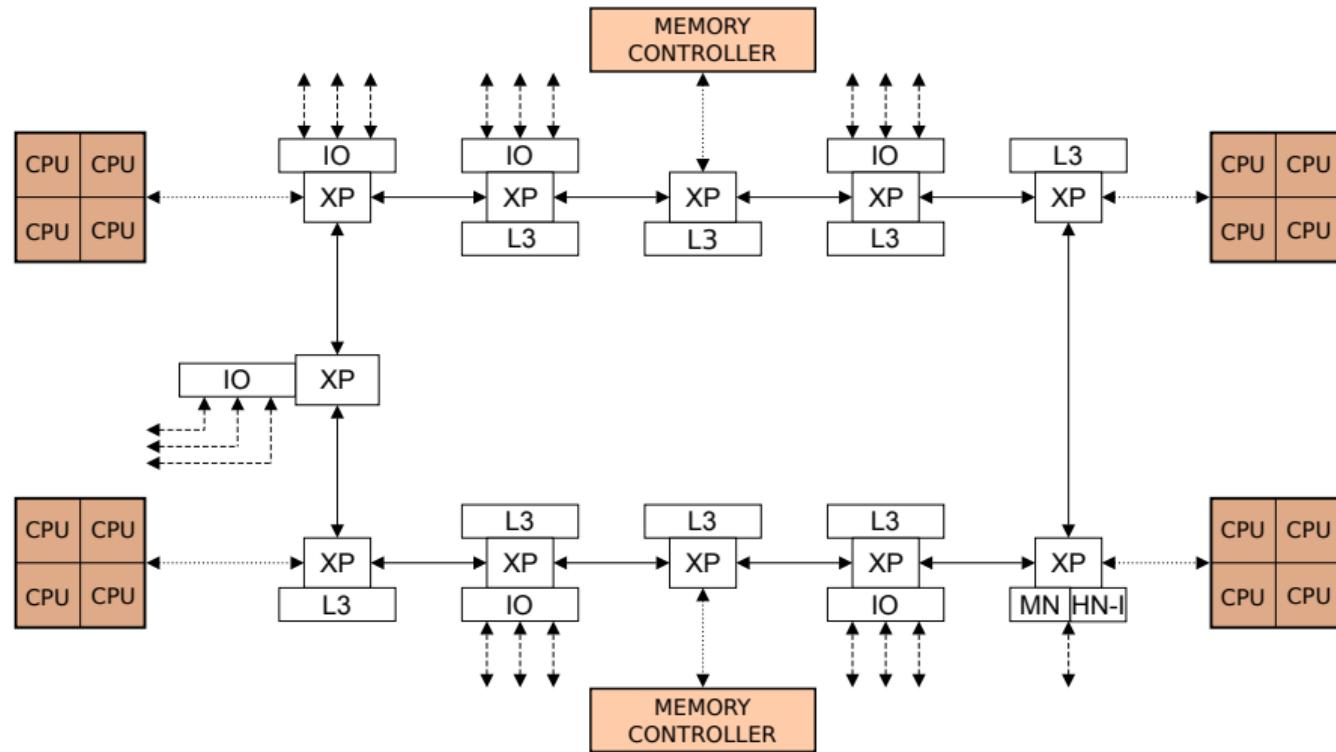
1	4952.50	14890.49
2	4955.55	14900.19
3	4949.13	14879.60
4	4954.66	14896.26

## perf and off- (or un-)core data

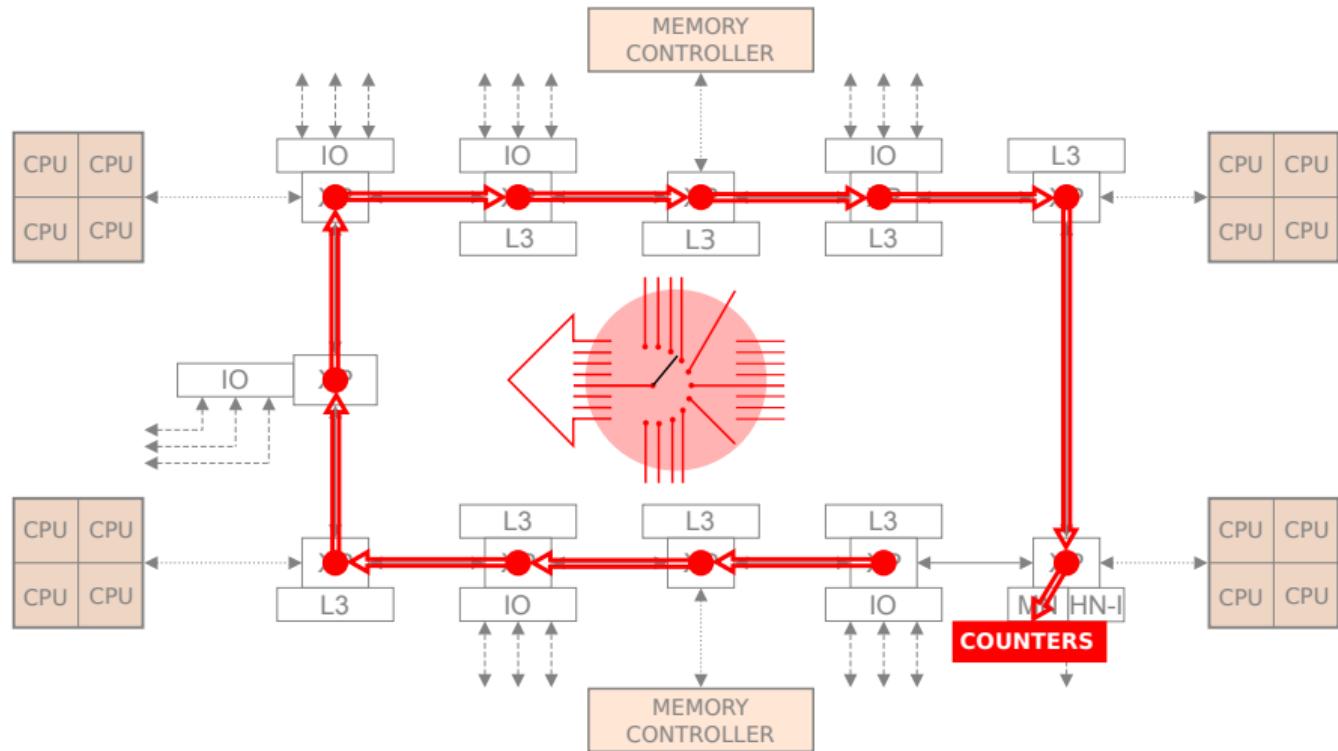
- -a
- -c 0
- /sys/bus/event\_source/devices/\*/cpumask
  - hotplug issues
  - “*perf\_pmu\_migrate\_context() is just f\*cked up. It needs to be reverted.*” (<https://lkml.org/lkml/2014/9/5/525>)

### Possible improvement:

- cpumask v2 - request only once, no CPU affinity, handle interrupt on any CPU



# CCN-504 Debug and Trace subsystem



## drivers/bus/arm-ccn.c

- perf's struct pmu
- Merged in v3.17
- DT + auto-discovery
  - no topology data

### Possible improvements:

- cpumask or equivalent
- events rotation
- documentation (or rather: more of it :-)
- topology

## CCN events

```
/ # perf list | grep ccn
ccn/cycles/                                [Kernel PMU event]
ccn/hnf_cache_fill/                          [Kernel PMU event]
ccn/hnf_cache_miss/                          [Kernel PMU event]
ccn/hnf_13_eviction/                         [Kernel PMU event]
ccn/hnf_13_fill_invalid_way/                 [Kernel PMU event]
ccn/hnf_13_sf_cache_access/                  [Kernel PMU event]
ccn/hnf_mc_reqs/                            [Kernel PMU event]
ccn/hnf_mc_retries/                          [Kernel PMU event]
ccn/hnf_pocq_reqs_recv/                     [Kernel PMU event]
ccn/hnf_pocq_retry/                           [Kernel PMU event]
ccn/hnf_qos_hh_retry/                         [Kernel PMU event]
ccn/hnf_sf_evictions/                        [Kernel PMU event]
```

## CCN events

ccn/hnf_sf_hit/	[Kernel PMU event]
ccn/hnf_snoops_broadcast/	[Kernel PMU event]
ccn/hnf_snoops_sent/	[Kernel PMU event]
ccn/hni_rxdatflits/	[Kernel PMU event]
ccn/hni_rxreqflits/	[Kernel PMU event]
ccn/hni_rxreqflits_order/	[Kernel PMU event]
ccn/hni_txdatflits/	[Kernel PMU event]
ccn/hni_txreqflits/	[Kernel PMU event]
ccn/mn_dvmop/	[Kernel PMU event]
ccn/mn_ecbarrier/	[Kernel PMU event]
ccn/mn_eobarrier/	[Kernel PMU event]
ccn/rni_rdata_beats_p0/	[Kernel PMU event]
ccn/rni_rdata_beats_p1/	[Kernel PMU event]

## CCN events

ccn/rni_rdata_beats_p2/	[Kernel PMU event]
ccn/rni_rrt_full/	[Kernel PMU event]
ccn/rni_rxdat_flits/	[Kernel PMU event]
ccn/rni_txdat_flits/	[Kernel PMU event]
ccn/rni_txreq_flits/	[Kernel PMU event]
ccn/rni_txreq_flits_replayed/	[Kernel PMU event]
ccn/rni_txreq_flits_retried/	[Kernel PMU event]
ccn/rni_wrt_full/	[Kernel PMU event]
ccn/sbas_rdata_beats_p0/	[Kernel PMU event]
ccn/sbas_rrt_full/	[Kernel PMU event]
ccn/sbas_rxdat_flits/	[Kernel PMU event]
ccn/sbas_txdat_flits/	[Kernel PMU event]
ccn/sbas_txreq_flits/	[Kernel PMU event]

## CCN events

ccn/sbas_txreq_flits_replayed	[Kernel PMU event]
ccn/sbas_txreq_flits_retried/	[Kernel PMU event]
ccn/sbas_wrt_full/	[Kernel PMU event]
ccn/sbsx_rxdatflits/	[Kernel PMU event]
ccn/sbsx_rxreqflits/	[Kernel PMU event]
ccn/sbsx_rxreqflits_order/	[Kernel PMU event]
ccn/sbsx_txdatflits/	[Kernel PMU event]
ccn/sbsx_txreqflits/	[Kernel PMU event]
ccn/xp_download_starvation/	[Kernel PMU event]
ccn/xp_respin/	[Kernel PMU event]
ccn/xp_upload_starvation/	[Kernel PMU event]
ccn/xp_valid_flit/	[Kernel PMU event]
ccn/xp_watchpoint/	[Kernel PMU event]

## Complex configuration

```
struct perf_event_attr {  
[...]  
    __u64 config;  
[...]  
    union {  
        __u64 bp_addr;  
        __u64 config1; /* extension of config */  
    };  
    union {  
        __u64 bp_len;  
        __u64 config2; /* extension of config1 */  
    };  
[...]
```

## Configuration format

```
/sys/bus/event_source/devices/ccn/format/node: config:0-7
/sys/bus/event_source/devices/ccn/format/xp: config:0-7
/sys/bus/event_source/devices/ccn/format/type: config:8-15
/sys/bus/event_source/devices/ccn/format/event: config:16-23
/sys/bus/event_source/devices/ccn/format/port: config:24-25
/sys/bus/event_source/devices/ccn/format/vc: config:26-28
/sys/bus/event_source/devices/ccn/format/dir: config:29-29
/sys/bus/event_source/devices/ccn/format/mask: config:30-33
/sys/bus/event_source/devices/ccn/format/cmp_l: config1:0-62
/sys/bus/event_source/devices/ccn/format/cmp_h: config2:0-59
```

# Configuration events

```
/sys/bus/event_source/devices/ccn/events/\  
    cycles:          type=0xff  
    hnf_cache_fill: type=0x4,event=0x3  
    hnf_cache_miss: type=0x4,event=0x1
```

- Equivalent commands:

```
# perf stat -C 0 -e ccn/hnf_cache_fill,node=3/ sleep 1  
# perf stat -C 0 -e ccn/type=0x4,event=0x3,node=3/ sleep 1  
# perf stat -C 0 -e ccn/config=0x030403/ sleep 1
```

## Possible improvement:

- Extended configuration scheme (strings?)

# Watchpoints

- two 64-bit values (plus two masks) comparator
- complex scenarios
- config[ 12] (3x64 bits) not enough
- sysfs cmp\_masks workaround

```
/sys/bus/event_source/devices/ccn/events/xp_watchpoint:  
    type=0x8, event=0xfe  
# perf stat -C 0 -e \  
    ccn/xp_watchpoint,dir=X,vc=X,cmp_h=X,cmp_l=X,mask=0xX/ ...
```

## Possible improvements:

- Extended configuration scheme (strings?)
- Support complex setups

`/sys/bus/event_source/devices/ccn/cmp_mask/*`

`/sys/bus/event_source/devices/ccn/cmp_mask/0h:0x0000000000000000`  
`/sys/bus/event_source/devices/ccn/cmp_mask/01:0x0000000000000000`  
[...]  
`/sys/bus/event_source/devices/ccn/cmp_mask/7h:0x0000000000000000`  
`/sys/bus/event_source/devices/ccn/cmp_mask/71:0x0000000000000000`  
`/sys/bus/event_source/devices/ccn/cmp_mask/8h:0xffffffffffffffff`  
`/sys/bus/event_source/devices/ccn/cmp_mask/81:0xffffffffffffffff`  
`/sys/bus/event_source/devices/ccn/cmp_mask/9h:0x0000000000000000`  
`/sys/bus/event_source/devices/ccn/cmp_mask/91:0x0000000000000000`  
`/sys/bus/event_source/devices/ccn/cmp_mask/ah:0xffffffffffff7fff`  
`/sys/bus/event_source/devices/ccn/cmp_mask/al:0xffffffffffffffff`  
`/sys/bus/event_source/devices/ccn/cmp_mask/bh:0xffffffffffffc1ff`  
`/sys/bus/event_source/devices/ccn/cmp_mask/bl:0xffffffffffffffff`

# Watchpoint-based events

- Pseudo-PMU events, defined as watchpoints

```
/sys/bus/event_source/devices/ccn/events/sbsx_rxreqflits_order:  
    type=0xc, event=0xfe, dir=0, vc=0, cmp_h=0x8000, mask=0xa  
/sys/bus/event_source/devices/ccn/events/mn_dvmop:  
    type=0x1, event=0xfe, dir=0, vc=0, cmp_h=0x2800, mask=0xb
```

## Possible improvement:

- BFP?

# Data sampling

- Selection of sampled values:
  - `PERF_SAMPLE_IP`
  - `PERF_SAMPLE_CALLCHAIN`
  - `PERF_SAMPLE_CPU`
  - `PERF_SAMPLE_RAW`
- No PC sampling for CCN
  - but periodic data still interesting
  - CPU/hrtimer driven groups

## Possible improvements:

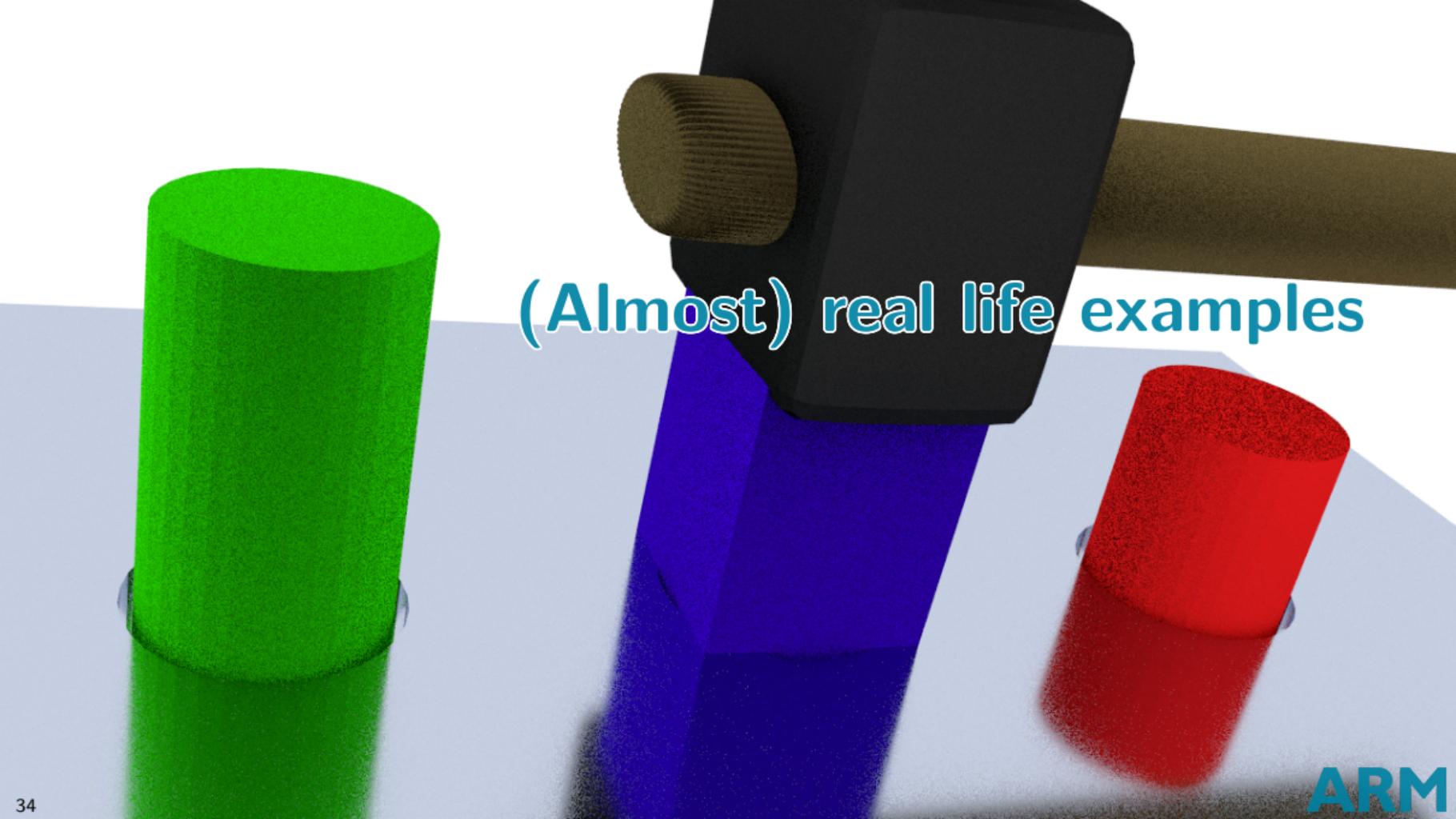
- CPU (bus master) sampling
- messages sampling
- support for grouping in the perf tool

# Complex metrics

- Cache miss ratio
  - ccn/hnf\_cache\_miss/
  - ccn/hnf\_l3\_sf\_cache\_access/
- Issues with rotation
  - Always rotating groups

## Possible improvements:

- user-space calculations (<http://lwn.net/Articles/532634/>)
- BPF?
- “recipes” (heuristic scripts)



(Almost) real life examples

## Idle system

```
# perf stat -C 0 -e ccn/cycles/, \
    /xp_valid_flit,xp=0,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=0,dir=1,vc=3/ sleep 12
```

Performance counter stats for 'CPU(s) 0':

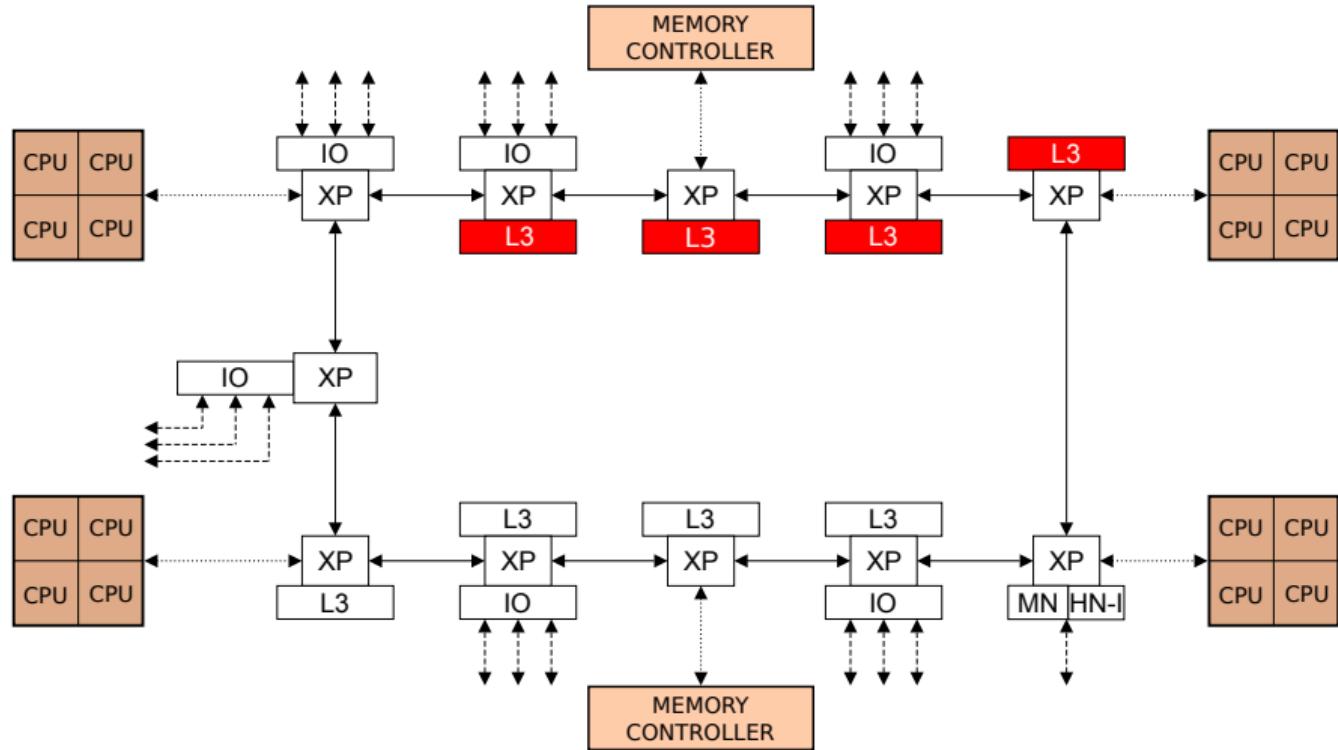
14401524916	ccn/cycles/
123137	ccn/xp_valid_flit,xp=0,dir=0,vc=3/
123128	ccn/xp_valid_flit,xp=0,dir=1,vc=3/
12.001203023	seconds time elapsed

- Bus utilisation:  $\frac{123137+123128}{14401524916} = 17.1 * 10^{-6}$

## Idle system

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=3/, \
    ccn/hnf_13_sf_cache_access,node=3/, \
    ccn/hnf_cache_miss,node=5/, \
    ccn/hnf_13_sf_cache_access,node=5/, \
    ccn/hnf_cache_miss,node=7/, \
    ccn/hnf_13_sf_cache_access,node=7/, \
    ccn/hnf_cache_miss,node=8/, \
    ccn/hnf_13_sf_cache_access,node=8/ \
sleep 12
```

# Idle system



## Idle system

Performance counter stats for 'CPU(s) 0':

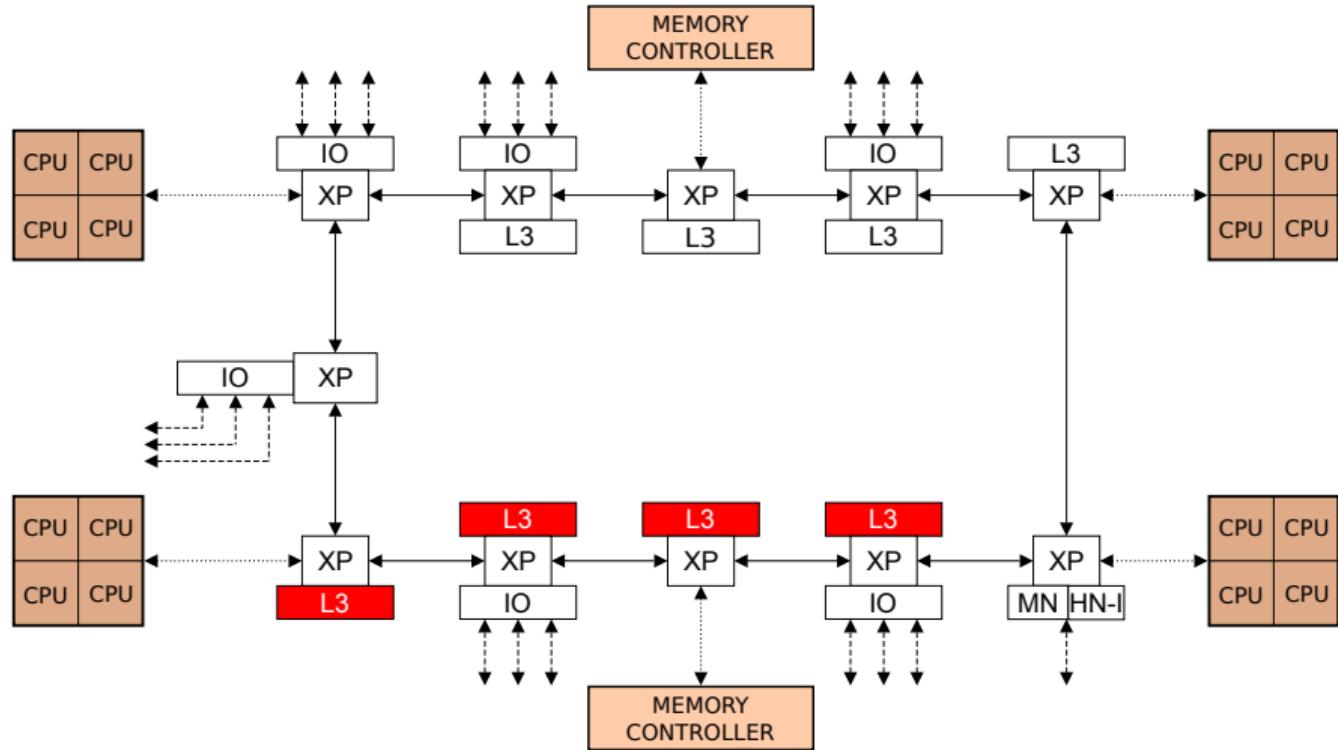
14402251731	ccn/cycles/
11338	ccn/hnf_cache_miss,node=3/
36735	ccn/hnf_13_sf_cache_access,node=3/
10395	ccn/hnf_cache_miss,node=5/
33698	ccn/hnf_13_sf_cache_access,node=5/
9206	ccn/hnf_cache_miss,node=7/
29222	ccn/hnf_13_sf_cache_access,node=7/
9765	ccn/hnf_cache_miss,node=8/
32207	ccn/hnf_13_sf_cache_access,node=8/
12.001506031	seconds time elapsed

- Cache miss ratios:  $\frac{11338}{36735} = 0.31$ ,  $\frac{10395}{33698} = 0.31$ ,  $\frac{9206}{29222} = 0.32$ ,  $\frac{9765}{32207} = 0.30$

## Idle system

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=13/, \
    ccn/hnf_13_sf_cache_access,node=13/, \
    ccn/hnf_cache_miss,node=15/, \
    ccn/hnf_13_sf_cache_access,node=15/, \
    ccn/hnf_cache_miss,node=17/, \
    ccn/hnf_13_sf_cache_access,node=17/, \
    ccn/hnf_cache_miss,node=18/, \
    ccn/hnf_13_sf_cache_access,node=18/ \
sleep 12
```

# Idle system



## Idle system

Performance counter stats for 'CPU(s) 0':

14402963391	ccn/cycles/
5198	ccn/hnf_cache_miss,node=13/
17035	ccn/hnf_13_sf_cache_access,node=13/
9901	ccn/hnf_cache_miss,node=15/
30448	ccn/hnf_13_sf_cache_access,node=15/
10908	ccn/hnf_cache_miss,node=17/
33404	ccn/hnf_13_sf_cache_access,node=17/
17984	ccn/hnf_cache_miss,node=18/
46203	ccn/hnf_13_sf_cache_access,node=18/
12.002100898	seconds time elapsed

- Cache miss ratios: 0.31, 0.33, 0.33, 0.39

## CPU intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/xp_valid_flit,xp=0,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=0,dir=1,vc=3/ \
    ./dhry.sh
```

Performance counter stats for 'CPU(s) 0':

15189653552	ccn/cycles/
133591	ccn/xp_valid_flit,xp=0,dir=0,vc=3/
133570	ccn/xp_valid_flit,xp=0,dir=1,vc=3/
12.657976711	seconds time elapsed

- Bus utilisation:  $17.6 \times 10^{-6}$
- Idle system was:  $17.1 \times 10^{-6}$

## CPU intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=3/, \
    ccn/hnf_13_sf_cache_access,node=3/, \
    ccn/hnf_cache_miss,node=5/, \
    ccn/hnf_13_sf_cache_access,node=5/, \
    ccn/hnf_cache_miss,node=7/, \
    ccn/hnf_13_sf_cache_access,node=7/, \
    ccn/hnf_cache_miss,node=8/, \
    ccn/hnf_13_sf_cache_access,node=8/ \
    ./dhry.sh
```

# CPU intensive

Performance counter stats for 'CPU(s) 0':

15191299475	ccn/cycles/
13061	ccn/hnf_cache_miss,node=3/
38663	ccn/hnf_13_sf_cache_access,node=3/
10435	ccn/hnf_cache_miss,node=5/
31699	ccn/hnf_13_sf_cache_access,node=5/
10018	ccn/hnf_cache_miss,node=7/
28670	ccn/hnf_13_sf_cache_access,node=7/
10009	ccn/hnf_cache_miss,node=8/
30803	ccn/hnf_13_sf_cache_access,node=8/
12.659046141	seconds time elapsed

- Cache miss ratios: 0.34, 0.33, 0.35, 0.32
- Idle system was: 0.31, 0.31, 0.32, 0.30

## CPU intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=13/, \
    ccn/hnf_13_sf_cache_access,node=13/, \
    ccn/hnf_cache_miss,node=15/, \
    ccn/hnf_13_sf_cache_access,node=15/, \
    ccn/hnf_cache_miss,node=17/, \
    ccn/hnf_13_sf_cache_access,node=17/, \
    ccn/hnf_cache_miss,node=18/, \
    ccn/hnf_13_sf_cache_access,node=18/ \
    ./dhry.sh
```

# CPU intensive

Performance counter stats for 'CPU(s) 0':

15199379301	ccn/cycles/
6438	ccn/hnf_cache_miss,node=13/
17447	ccn/hnf_13_sf_cache_access,node=13/
9485	ccn/hnf_cache_miss,node=15/
25009	ccn/hnf_13_sf_cache_access,node=15/
11248	ccn/hnf_cache_miss,node=17/
30264	ccn/hnf_13_sf_cache_access,node=17/
19256	ccn/hnf_cache_miss,node=18/
44188	ccn/hnf_13_sf_cache_access,node=18/
12.665779109	seconds time elapsed

- Cache miss ratios: 0.37, 0.38, 0.37, 0.44
- Idle system was: 0.31, 0.33, 0.33, 0.39

## Memory intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/xp_valid_flit,xp=0,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=0,dir=1,vc=3/ \
    ./stream.sh
```

Performance counter stats for 'CPU(s) 0':

18179667260	ccn/cycles/
626886869	ccn/xp_valid_flit,xp=0,dir=0,vc=3/
626886516	ccn/xp_valid_flit,xp=0,dir=1,vc=3/
15.149639312	seconds time elapsed

- Bus utilisation:  $68965.7 \times 10^{-6}$
- CPU intensive was:  $17.6 \times 10^{-6}$  (3919x!)
- Idle system was:  $17.1 \times 10^{-6}$

## Memory intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=3/, \
    ccn/hnf_13_sf_cache_access,node=3/, \
    ccn/hnf_cache_miss,node=5/, \
    ccn/hnf_13_sf_cache_access,node=5/, \
    ccn/hnf_cache_miss,node=7/, \
    ccn/hnf_13_sf_cache_access,node=7/, \
    ccn/hnf_cache_miss,node=8/, \
    ccn/hnf_13_sf_cache_access,node=8/ \
    ./stream.sh
```

# Memory intensive

Performance counter stats for 'CPU(s) 0':

18183426085	ccn/cycles/
37537969	ccn/hnf_cache_miss,node=3/
38011948	ccn/hnf_13_sf_cache_access,node=3/
37533738	ccn/hnf_cache_miss,node=5/
38004426	ccn/hnf_13_sf_cache_access,node=5/
37528567	ccn/hnf_cache_miss,node=7/
37994670	ccn/hnf_13_sf_cache_access,node=7/
37539331	ccn/hnf_cache_miss,node=8/
38007115	ccn/hnf_13_sf_cache_access,node=8/
15.152468367	seconds time elapsed

- Cache miss ratios: 0.99, 0.99, 0.99, 0.99
- CPU intensive was: 0.34, 0.33, 0.35, 0.32
- Idle system was: 0.31, 0.31, 0.32, 0.30

## Memory intensive

```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/hnf_cache_miss,node=13/, \
    ccn/hnf_13_sf_cache_access,node=13/, \
    ccn/hnf_cache_miss,node=15/, \
    ccn/hnf_13_sf_cache_access,node=15/, \
    ccn/hnf_cache_miss,node=17/, \
    ccn/hnf_13_sf_cache_access,node=17/, \
    ccn/hnf_cache_miss,node=18/, \
    ccn/hnf_13_sf_cache_access,node=18/ \
    ./stream.sh
```

# Memory intensive

Performance counter stats for 'CPU(s) 0':

18184672598	ccn/cycles/
37546905	ccn/hnf_cache_miss,node=13/
38006466	ccn/hnf_13_sf_cache_access,node=13/
37552602	ccn/hnf_cache_miss,node=15/
38013475	ccn/hnf_13_sf_cache_access,node=15/
37556325	ccn/hnf_cache_miss,node=17/
38024197	ccn/hnf_13_sf_cache_access,node=17/
37566370	ccn/hnf_cache_miss,node=18/
38037931	ccn/hnf_13_sf_cache_access,node=18/

15.153509718 seconds time elapsed

- Cache miss ratios: 0.99, 0.99, 0.99, 0.99
- CPU intensive was: 0.37, 0.38, 0.37, 0.44
- Idle system was: 0.31, 0.33, 0.33, 0.39

# Traffic routing

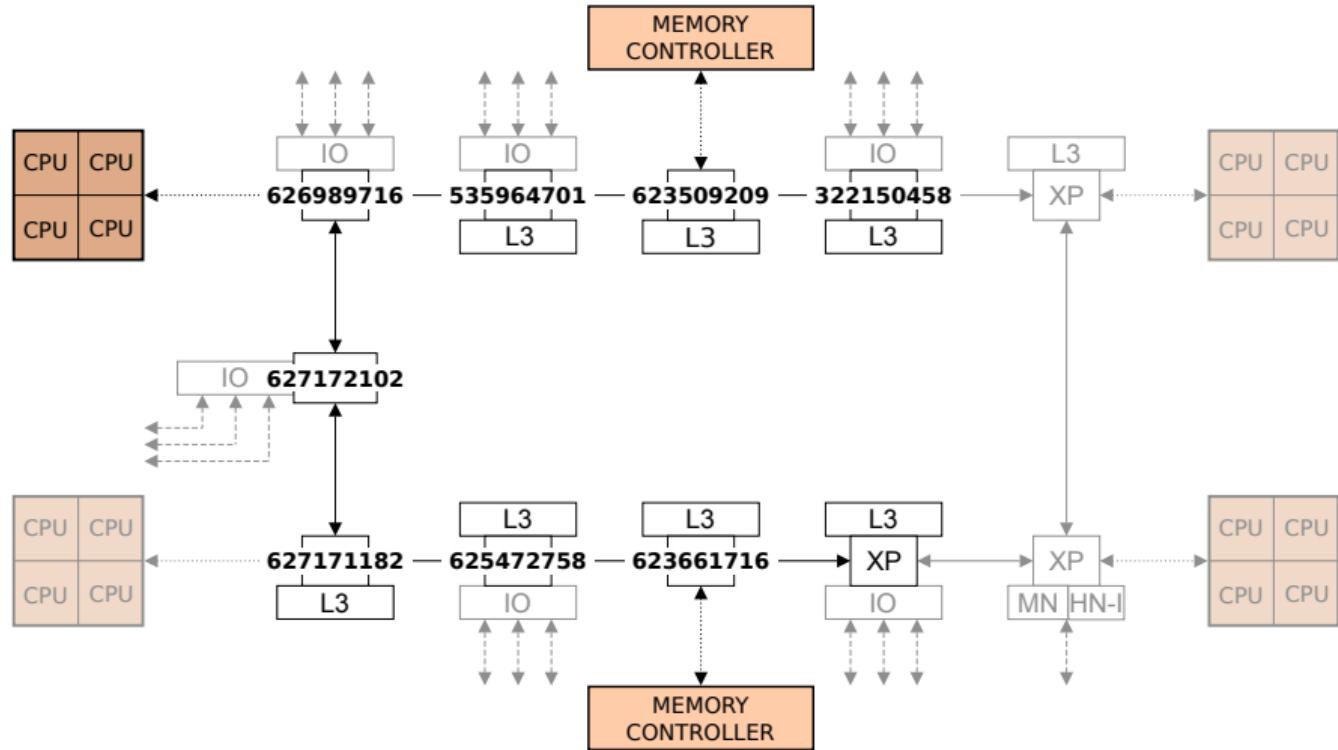
```
# perf stat -C 0 -e ccn/cycles/, \
    ccn/xp_valid_flit,xp=0,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=1,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=2,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=3,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=10,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=9,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=8,dir=0,vc=3/, \
    ccn/xp_valid_flit,xp=7,dir=0,vc=3/ \
    ./stream.sh
```

# Traffic routing

Performance counter stats for 'CPU(s) 0':

18181466817	ccn/cycles/
626989716	ccn/xp_valid_flit,xp=0,dir=0,vc=3/
535964701	ccn/xp_valid_flit,xp=1,dir=0,vc=3/
623509209	ccn/xp_valid_flit,xp=2,dir=0,vc=3/
322150458	ccn/xp_valid_flit,xp=3,dir=0,vc=3/
627172102	ccn/xp_valid_flit,xp=10,dir=0,vc=3/
627171182	ccn/xp_valid_flit,xp=9,dir=0,vc=3/
625472758	ccn/xp_valid_flit,xp=8,dir=0,vc=3/
623661716	ccn/xp_valid_flit,xp=7,dir=0,vc=3/

# Idle system



# Stay tuned!

- ARM Juno System Profiler
  - Interconnect bandwidth and latency monitors
- Intel Cache Monitoring
  - Cache occupancy, even on per-task basis
- And even more of this in the near future

# Thank You

*The trademarks featured in this presentation are registered and/or unregistered trademarks of ARM Limited (or its subsidiaries) in the EU and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.*