

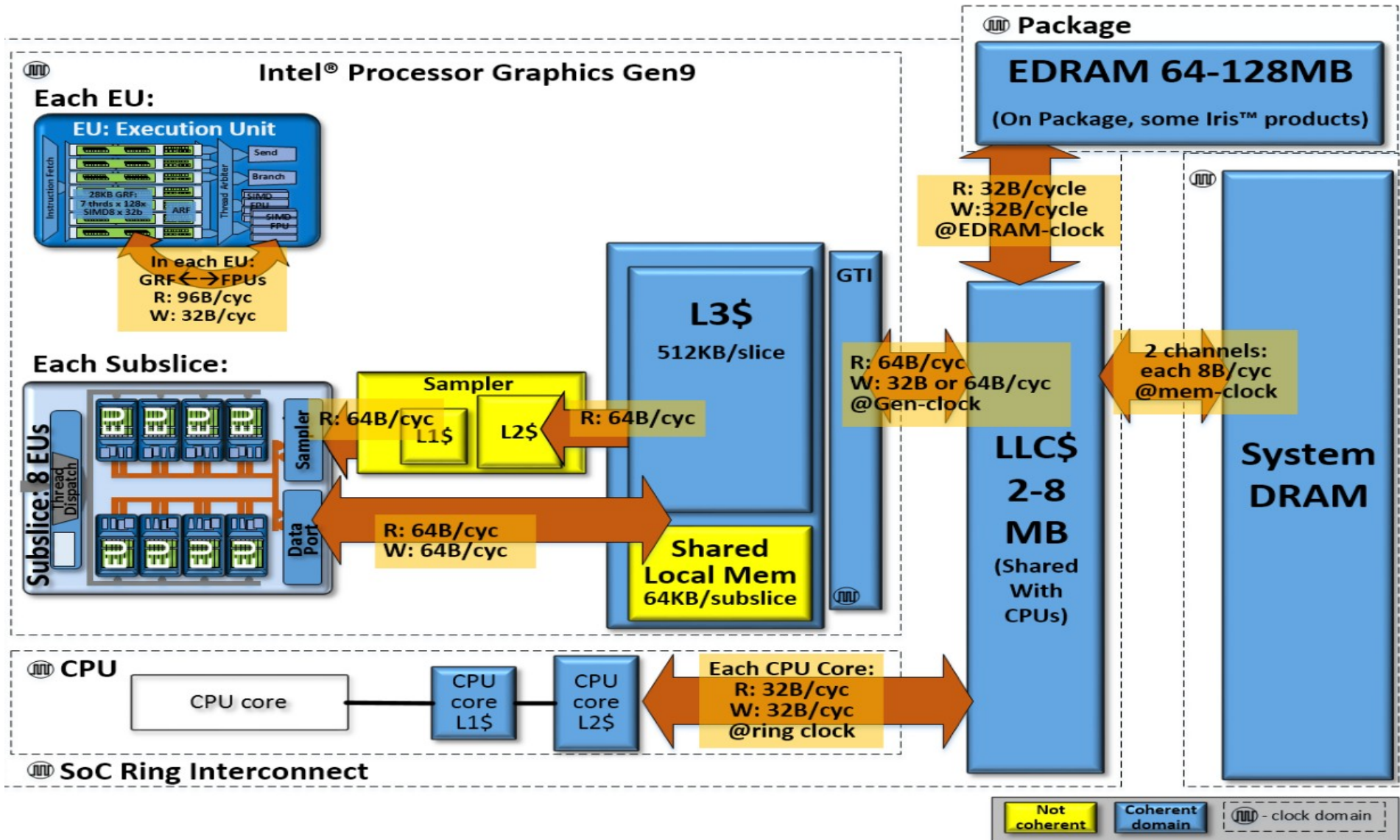
Realtime and Graphics – a Contradiction in Terms?

electronic displays Conference 2021
01. - 05. March 2021

Poster Session

Ahmed S. Darwish <a.darwish@linutronix.de>

Problem: GPUs share resources with CPUs



Source: "The Compute Architecture of Intel Processor Graphics Gen9", Intel

Noisy Neighbours: even with VM Isolation

The screenshot shows a Linux desktop environment with three main windows:

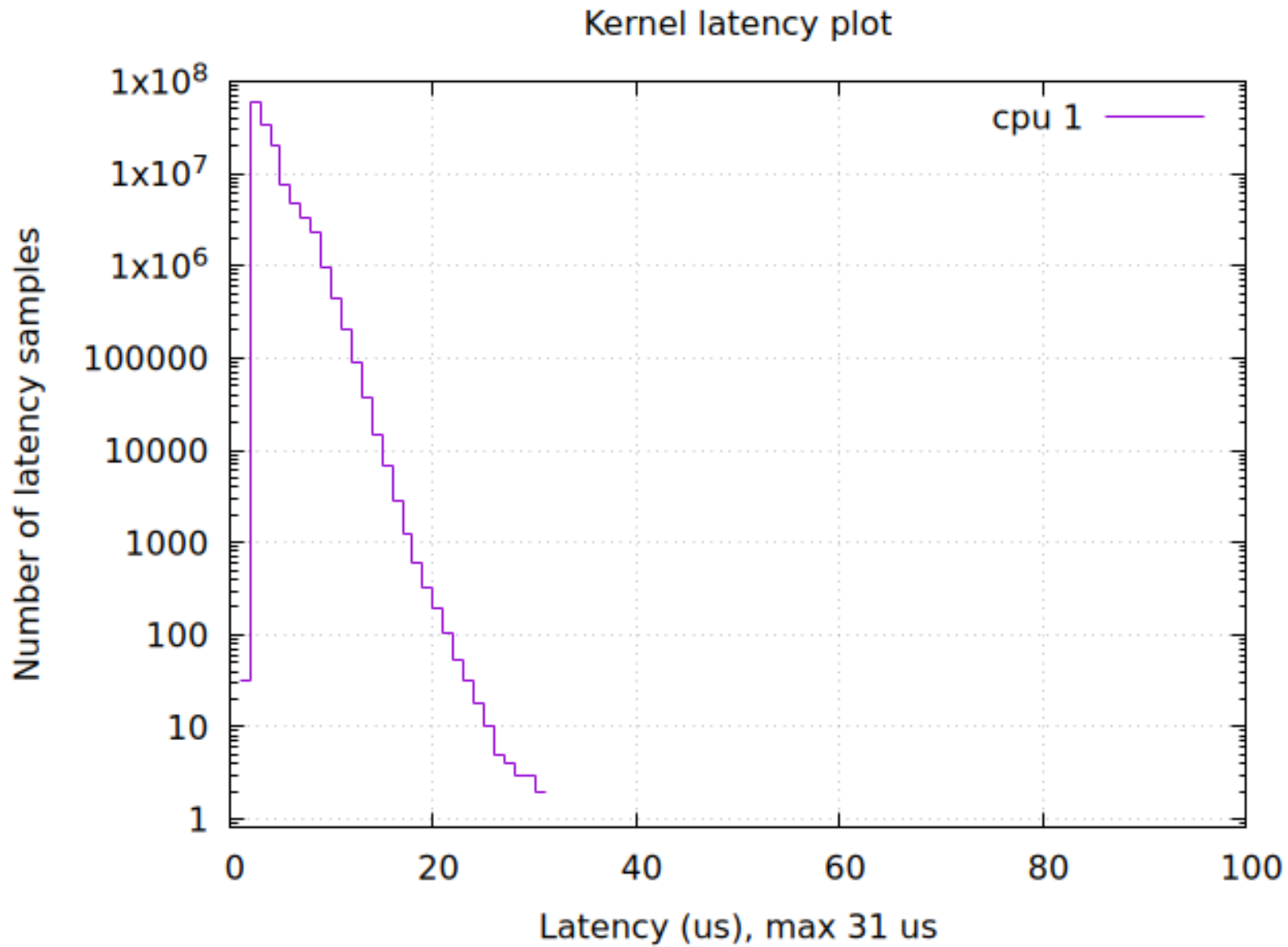
- Top Left:** A terminal window displaying a 3D rendered rabbit model. The terminal prompt is `darwi@rtvm:~$`.
- Top Right:** A terminal window showing GPU usage statistics. The output includes:


```
render busy: 100% ██████████ render space: 310/16384
task percent busy
CS: 100% ██████████ vert fetch: 14748482715 (35932371/sec)
GAM: 99% ██████████ prim fetch: 4916160985 (11977451/sec)
VS: 96% ██████████ VS invocations: 4965352930 (12097299/sec)
VF: 95% ██████████ GS invocations: 0 (0/sec)
CL: 95% ██████████ GS prims: 0 (0/sec)
GAFS: 90% ██████████ CL invocations: 4916160961 (11977393/sec)
SF: 89% ██████████ CL prims: 4605247075 (11172601/sec)
SDE: 21% ██████████ PS invocations: 91961835236 (224377692/sec)
GAFM: 9% ██████████ PS depth pass: 58822963751 (143622469/sec)
TDG: 2% ██████████
URBM: 1% ██████████
SVG: 0% ██████████
TSG: 0% ██████████
VFE: 0% ██████████
```
- Bottom Left:** A terminal window showing the execution of `sudo cyclictest` with various options. The output includes:

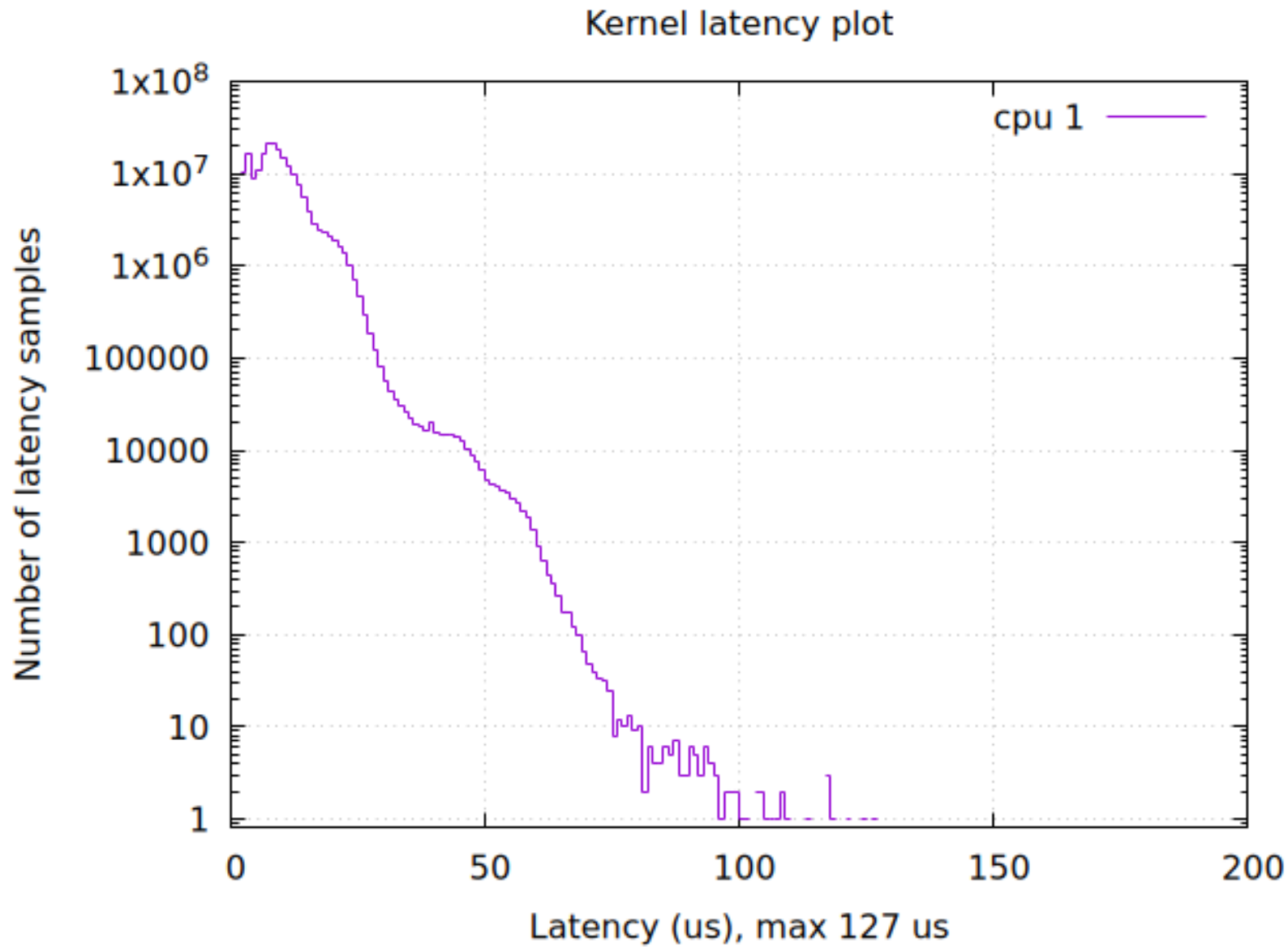

```
[sudo] password for darwi:
# /dev/cpu_dma_latency set to 0us
policy: fifo: loadavg: 0.42 0.48 0.52 1/169 844 policy: fifo: loadavg: 0.49
policy: fifo: loadavg: 0.57 0.52 0.54 1/169 844
policy: fifo: loadavg: 0.21 0.40 0.49 1/169 845
T: 0 ( 819) P:98 I:1000 C:3166400 Min: 3 Act: 5 Avg: 8 Max: 78
T: 0 ( 819) P:98 I:1000 C:3292492 Min: 3 Act: 11 Avg: 8 Max: 78
```
- Bottom Right:** A terminal window displaying a 3D rendered horse model.

Bottom Left: RT VM running over the ACRN Hypervisor + `cyclictest` latency measurement. Top right: GPU usage. Note: Graphical applications are running outside the RT virtual machine.

RT VM Latency: No External GPU Workload



RT VM Latency: with External GPU Workload



Segmenting Caches: Intel Cache Allocation Tech

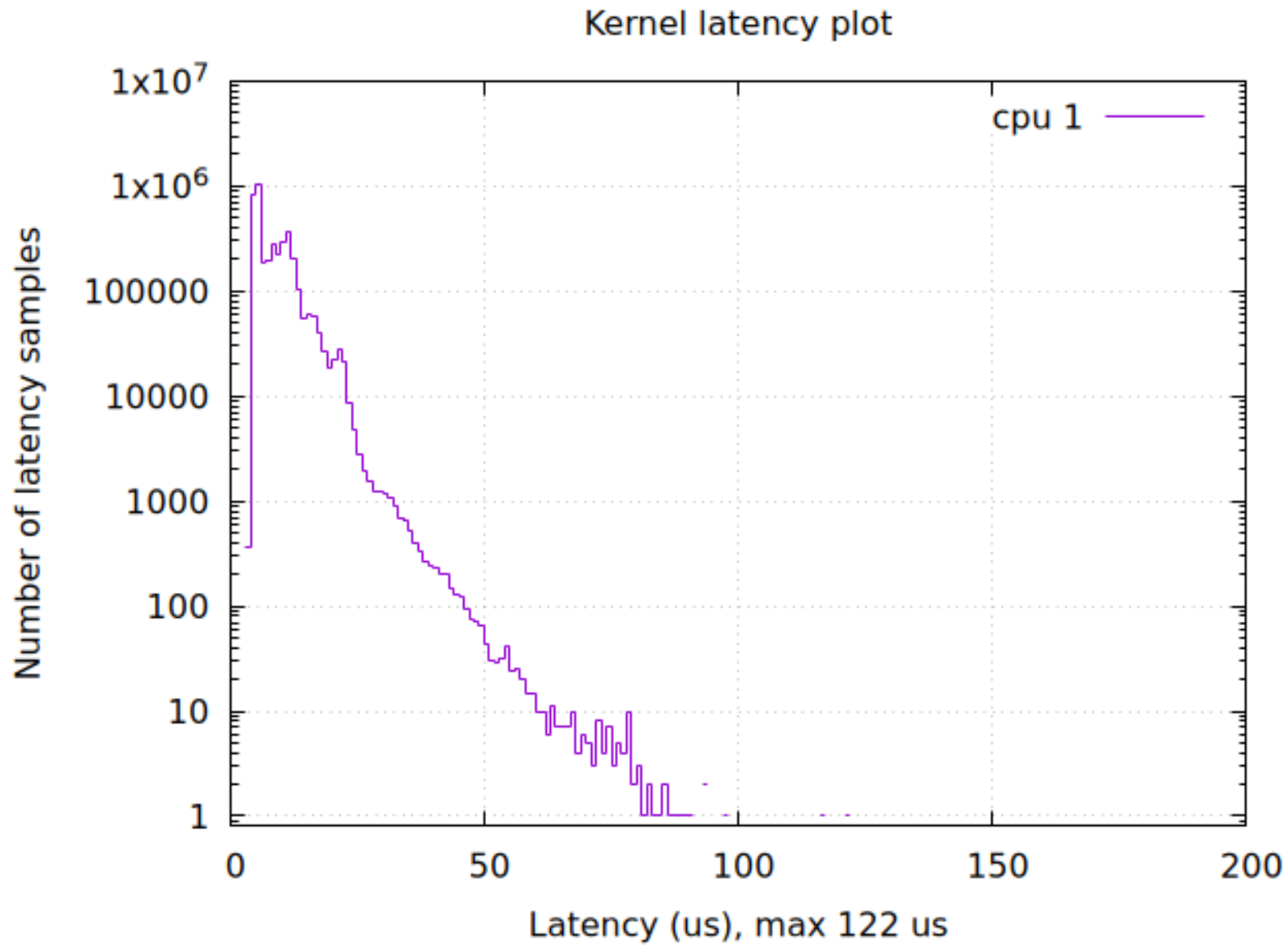
	M7	M6	M5	M4	M3	M2	M1	M0	
COS0	A	A	A	A	A	A	A	A	Default Bitmask
COS1	A	A	A	A	A	A	A	A	
COS2	A	A	A	A	A	A	A	A	
COS3	A	A	A	A	A	A	A	A	

	M7	M6	M5	M4	M3	M2	M1	M0	
COS0	A	A	A	A	A	A	A	A	Overlapped Bitmask
COS1					A	A	A	A	
COS2							A	A	
COS3								A	

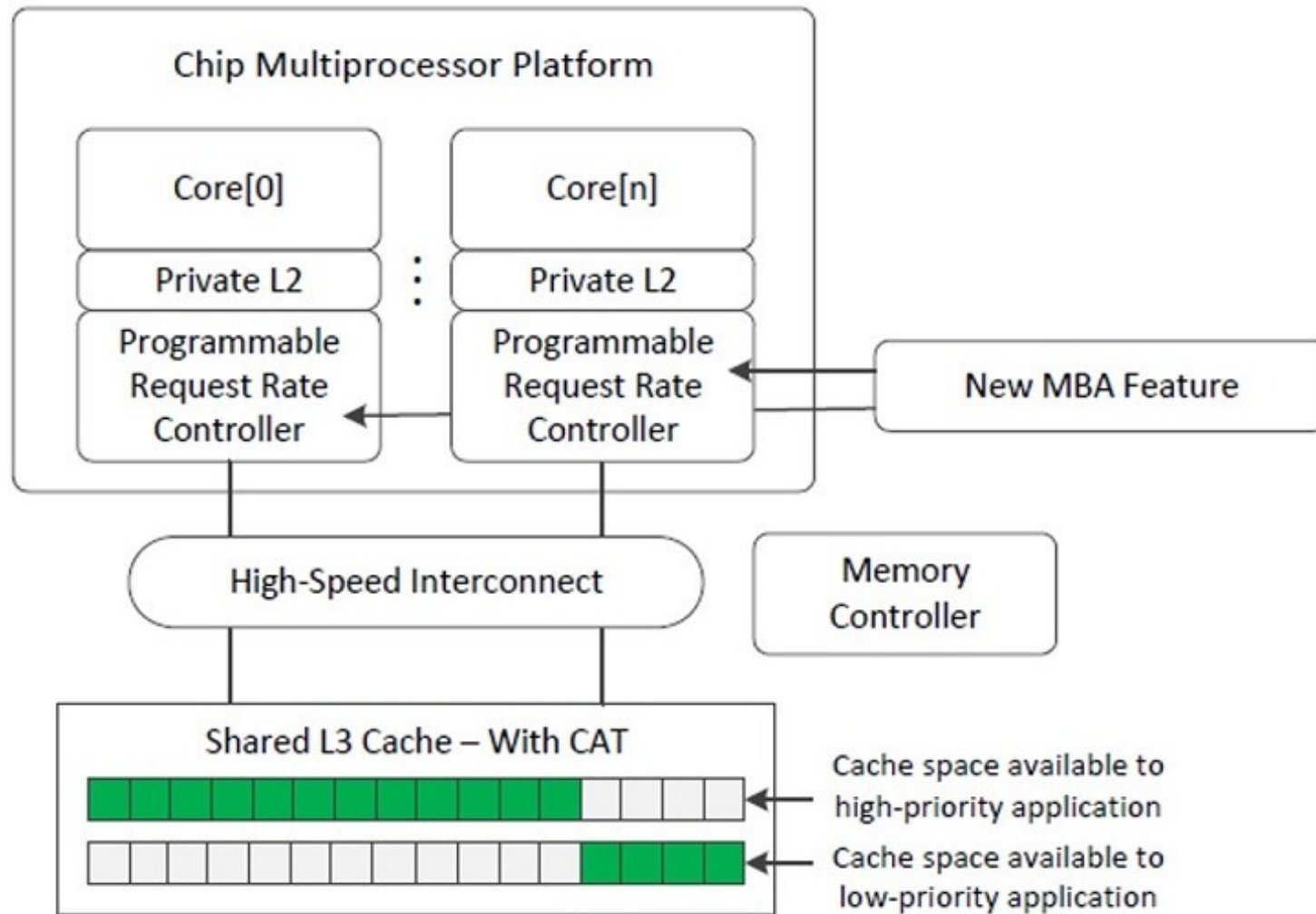
	M7	M6	M5	M4	M3	M2	M1	M0	
COS0	A	A	A	A					Isolated Bitmask
COS1					A	A			
COS2							A		
COS3								A	

Source: Intel® 64 and IA-32 Architectures Software Developer's Manual. Vol. 3A. Section 17.19.2 Cache Allocation Technology Architecture

RT VM Latency: External GPU Workload + CAT



Segmenting Memory Bandwidth: Intel MBA



Source: Intel® 64 and IA-32 Architectures Software Developer's Manual. Vol. 3A. Section 17.19.7 Introduction to Memory Bandwidth Allocation

Questions / Comments

Thank you for your attention.

a.darwish@linutronix.de

info@linutronix.de