A cycle-accurate methodology to improve PREM-like memory bandwidth underutilization on FPGA-based HeSoCs

Gianluca Brilli*, Giacomo Valente†, Alessandro Capotondi*,Tania di Mascio†, Paolo Burgio*, Paolo Valente* and Andrea Marongiu*

IWES, 2021

*__University of Modena and Reggio Emilia__, *<name>.<surname>@unimore.it*

†***University of L'Aquila***, *<name>.<surname>@univaq.it*

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

Fondo di Ateneo per la Ricerca FAR2020

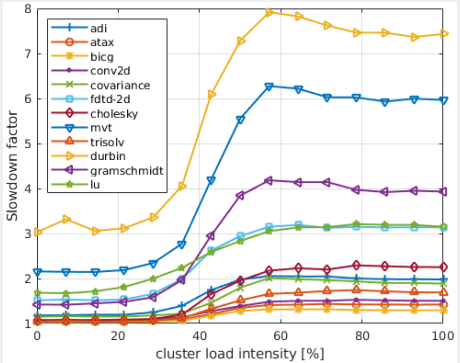
FRACTAL EDGE

COMP4DRONES

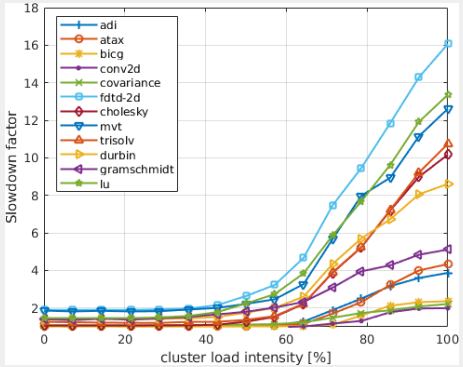
Motivations (1)

- As the number of engine grows on next generation of HeSoCs, the **interference** due to **shared interconnects** and **main memory** hampers tasks' execution time.

ZUS+ MPSoC (up to 8x)



Versal ACAP (up to 16x)



G. Brilli, A. Capotondi, P. Burgio and A. Marongiu, ***Understanding and Mitigating Memory Interference in FPGA-based HeSoCs***, 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2022.

Motivations (2)

- Available **memory bandwidth regulation** mechanisms are:
  - **Too Loosely-coupled** and **Coarse-grained** from the actuation & monitoring point of view**;**
  - or are **platform-specific.**

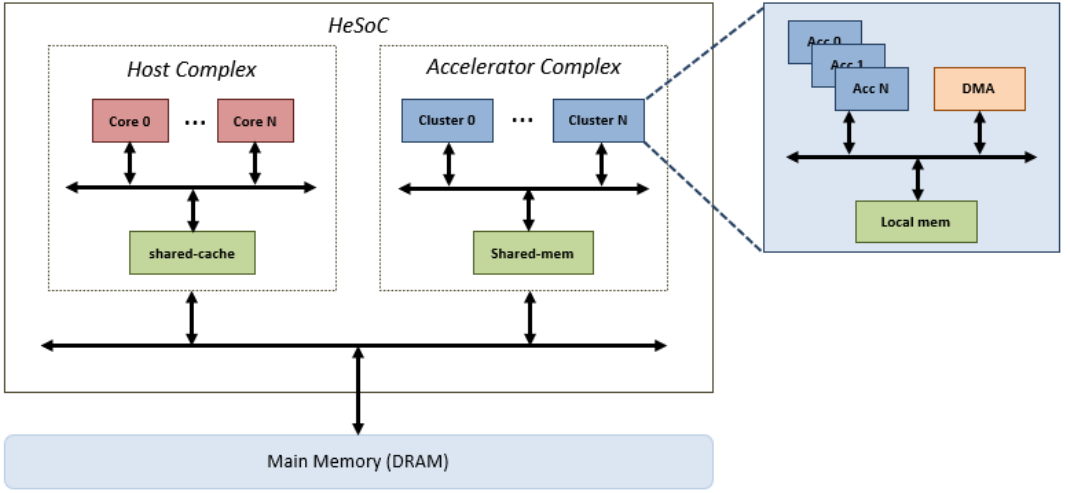
Contributions

- Runtime Bandwidth Regulator (RBR)
  - Tightly-coupled monitoring & throttling;
  - Minimal timing overhead (1 clock);
  - High precision QoS regulation.

  - Evaluation on Xilinx Zynq UltraScale+ MPSoC

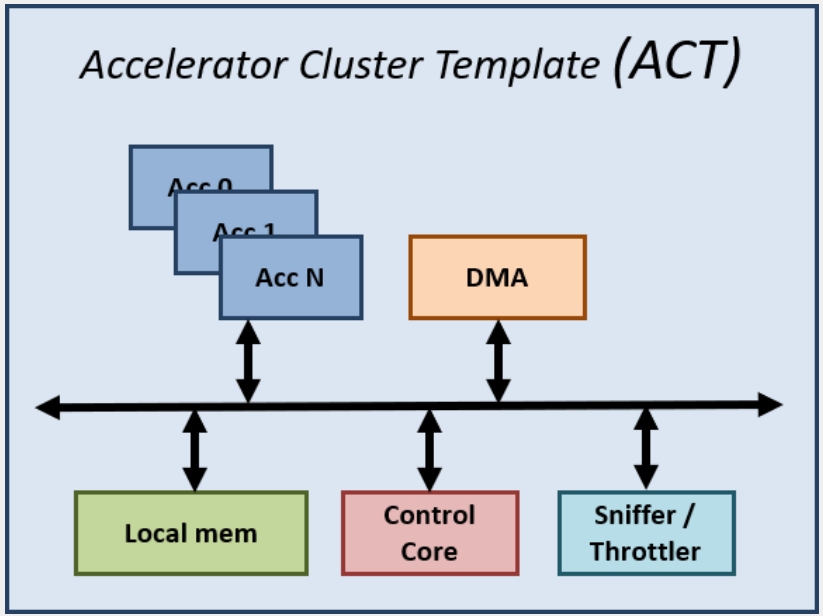
- **This work is currently under-submission! Can’t be disclosed.**

Background



- Some examples:
  - NVIDIA Xavier;
  - **Xilinx Zynq UltraSCALE+;**
  - Xilinx Versal.
- HeSoC is a **new emerging trend**.

The proposed mechanism



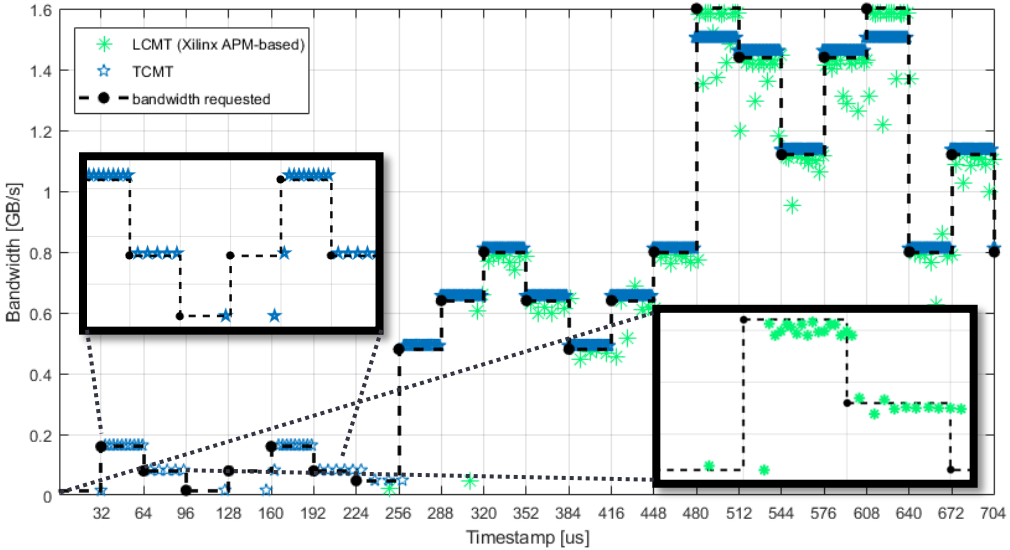
- **Control Core**: Tightly-coupled with DMA and Sniffer/Throttler.
  - Set the amount of bandwidth the DMA can use (CLI).
- **DMA**: performs **controlled** memory transactions.
- **Sniffer/Throttler**: constantly monitors the DMA activity and regulates DMA transactions.

Experimental Results (1)

- **Exp1** – Tightly-coupled versus Loosely-coupled Monitoring and Throttling.
  - **Objective**:
    - Test the ability of our system to follow a **bandwidth profile** (eg. provided by a system scheduler).
    - Compared with Loosely-coupled solutions on a Xilinx Zynq UltraScale+ MPSoC
      - based on Xilinx AXI Performance Monitor (APM).

Experimental Results (2)

- **Exp1** – **Tightly-coupled** versus **Loosely-coupled** Monitoring and Throttling.



- **Black dashed line**: bandwidth profile (e.g. system scheduler);
- **Blue**: our TC solution.
- **Green**: LC solution based on APM.

Experimental Results (3)

- **Exp1** – Tightly-coupled versus Loosely-coupled Monitoring and Throttling.
  - **Results**:
    - **Our Tightly-Coupled solution (TCMT),** follows a bandwidth profile with **32µs** of period;
    - **Platform-dependent Loosely-Coupled solutions (LCMT),** need a slower scheduling tick, at least **384µs** of period.
    - **12x** of improvement compared to Zynq UltraScale+ solutions.

Experimental Results (4)

- **Exp2** – QoS for Memory Interference Mitigation.

  - **Objective**:
    - Test the ability of our system to **mitigate memory interference** on Heterogeneous System (Xilinx ZUS+);

Experimental Results (5)

- **Exp2** – QoS for Memory Interference Mitigation.
  - **Exp-Setup**:
    - 3 ACT performing **controlled** memory reads;
    - Real applications on APU & RPU.
    - All the actors must meet deadlines (except ACT3 which is Best Effort)
  - **Scenarios**:
    - Very Tight (VT): max **20%** of tolerated slowdown
    - Tight (T): max **40%** of tolerated slowdown
    - Medium (M): max **60%** of tolerated slowdown

Experimental Results (6)

- **Exp2** – QoS for Memory Interference Mitigation.

Unfeasible configurations!

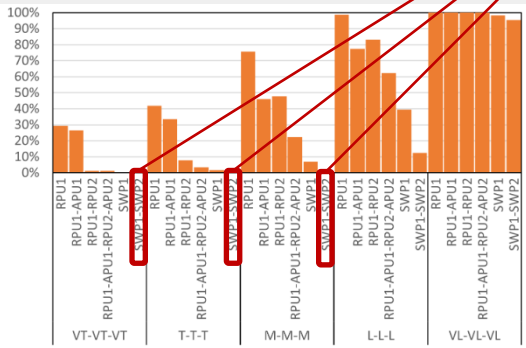


Figure 11 Ratio of accepted QoS setups with uniform thresholds for Workload 1.

- With ZUS+ QoS ecosystem.

Serrano-Cases, Alejandro, Juan M. Reina, Jaume Abella, Enrico Mezzetti and Francisco J. Cazorla. ***Leveraging Hardware QoS to Control Contention in the Xilinx Zynq UltraScale+ MPSoC***, ECRTS 2021.

Experimental Results (7)

- **Exp2** – QoS for Memory Interference Mitigation.

TABLE 4.2: Real-world benchmarks

Scenario	FPGA		APU		RPU	
	ACT1	ACT2	MM	MT	VMA	I2C
VT	1.20×	1.20×	1.22×	1.07×	1.70×	1.11×
	1.20×	-	1.22×	1.07×	1.22×	1.02×
	-	-	1.22×	1.07×	1.12×	1.01×
T	1.38×	1.38×	1.22×	1.07×	1.35×	1.04×
M	1.59×	1.59×	1.22×	1.07×	1.31×	1.04×

- With our methodology.

Unfeasible configuration!

Feasible configurations!

Conclusion

- We introduced a **fine-grained QoS control** via **tightly-coupled bandwidth monitoring and regulation**.
  - 12x faster than loosely-coupled bandwidth regulation mechanisms of the Zynq UltraScale+ MPSoC;
  - Our mechanism is more accurate than ZUS+ based QoS ecosystem.

Thank you!

Gianluca Brilli