Background & Novelty
On-chip memory pursues many desirable characteristics such as high throughput, low latency, and low cost. However, without functional correctness as a prerequisite, these characteristics are merely illusory. This work:
1. The 1st open-source microarchitectural solution for GPGPU cache that take consistency and coherence into design consideration.
2. Leverages the RISC-V CPU-centric consistency model RVWMO to GPGPU.
3. Leverages Release-Consistency-Directed Coherence (RCDC) to provide coherence functionality, reducing both the burden on programming frameworks and the complexity of hardware implementation.

Intro to Memory Consistency & GPU Cache Coherence
In multi-core/multi-thread scenarios, the consistency problem can arise from out-of-order (OoO) execution. Different consistency models with varying degrees of strictness can lead to different valid execution results for the same multi-threaded program.

Invalidate data
M/H
Req
H

Access different cache line
data
1,2

Area

Flush

Invalidate all dirty cache line
F

Microarch behavior

Microarch behavior

Global Invalidate

Global Invalidate

Drain MSHR

Drain MSHR

Global Invalidate

Global Invalidate

Attached RCC

Coherence Function(s)

"Acquire" and "Release"

"Acquire" and "Release"

"Acquire" and "Release"

None

"Acquire" and "Release"

"Release"

Comparison & Result
In the table below, the number on the left of each column indicates the typical delay cycles for the corresponding request type when there is no blocking in the corresponding signal path. The numbers on the right of each column indicate potential blocking scenarios.

Implementation of Axiomatic RVWMO in Micro Arch
Release Consistency provides "Acquire" or "Release" semantics to indicate its constraint. In Release-Consistency-Directed Coherence, the combination of "Release" with dirty cache line flush and "Acquire" with L1 global invalidation satisfies the coherence requirement.

PPO Rule 4-7: For RVWMO consistency operations(RV32i FENCE, RV32a. aq & rl identifier), mapping them to corresponding microarchitectural behavior.