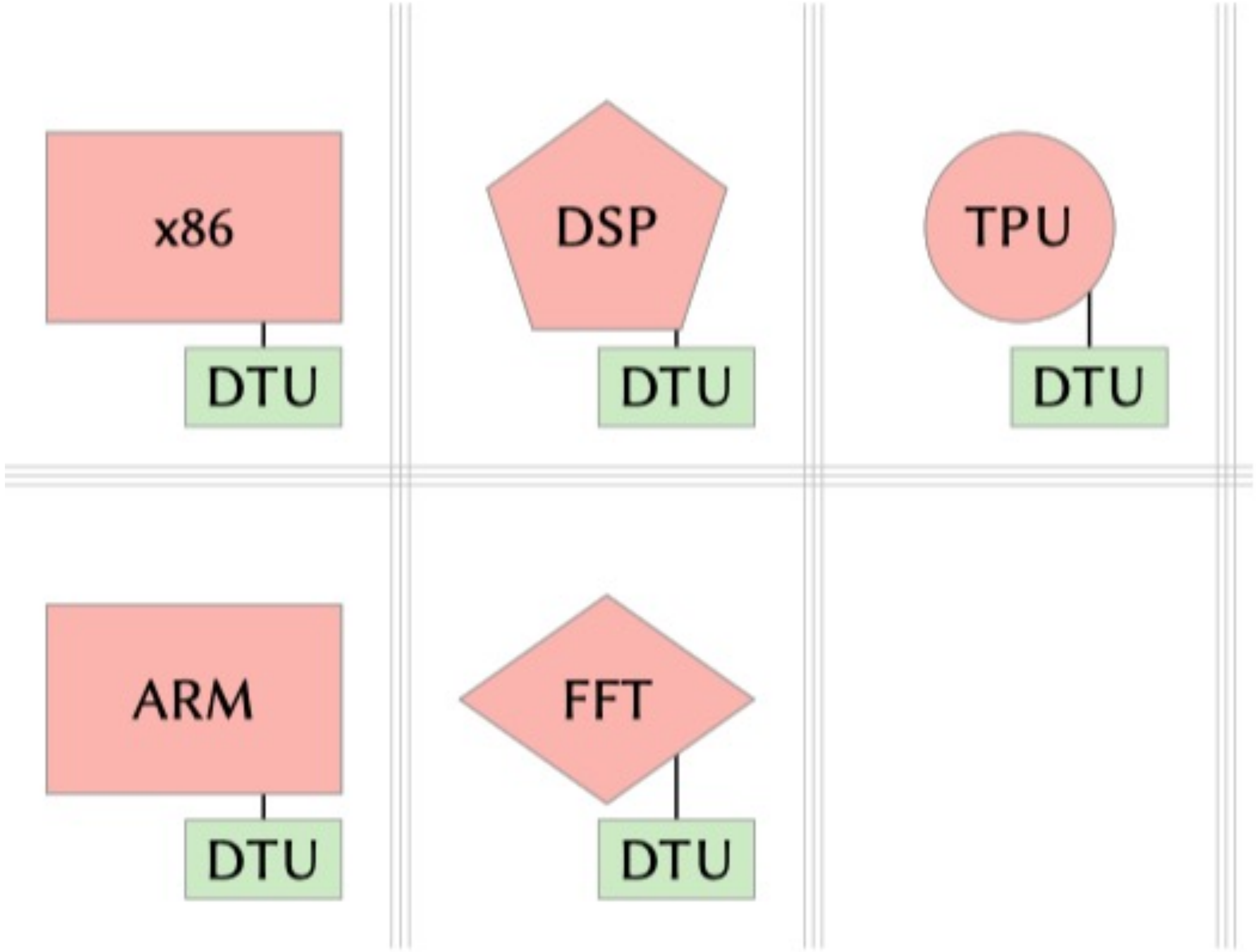
# M³x: Autonomous Accelerators via Context-Enabled Fast-Path Communication

# Contributions

- Rethink system architecture based on M$^3$
  - Hardware/operating-system co-design for heterogeneous systems
  - Simulation based on gem5
- Not built upon cache coherency
  - Costs (area, power, complexity, performance) increases with system size
  - More challenging for heterogeneous systems
  - Unclear whether future systems will be (globally) coherent
- Focus on fixed-function accelerators
  - Most difficult to support as "first-class citizens"
  - Provide none of the features OSes need

# Outline

- System Architecture and Background
- Autonomous Stream-Processing Accelerators
- Fast-Path Communication vs. Context Switching

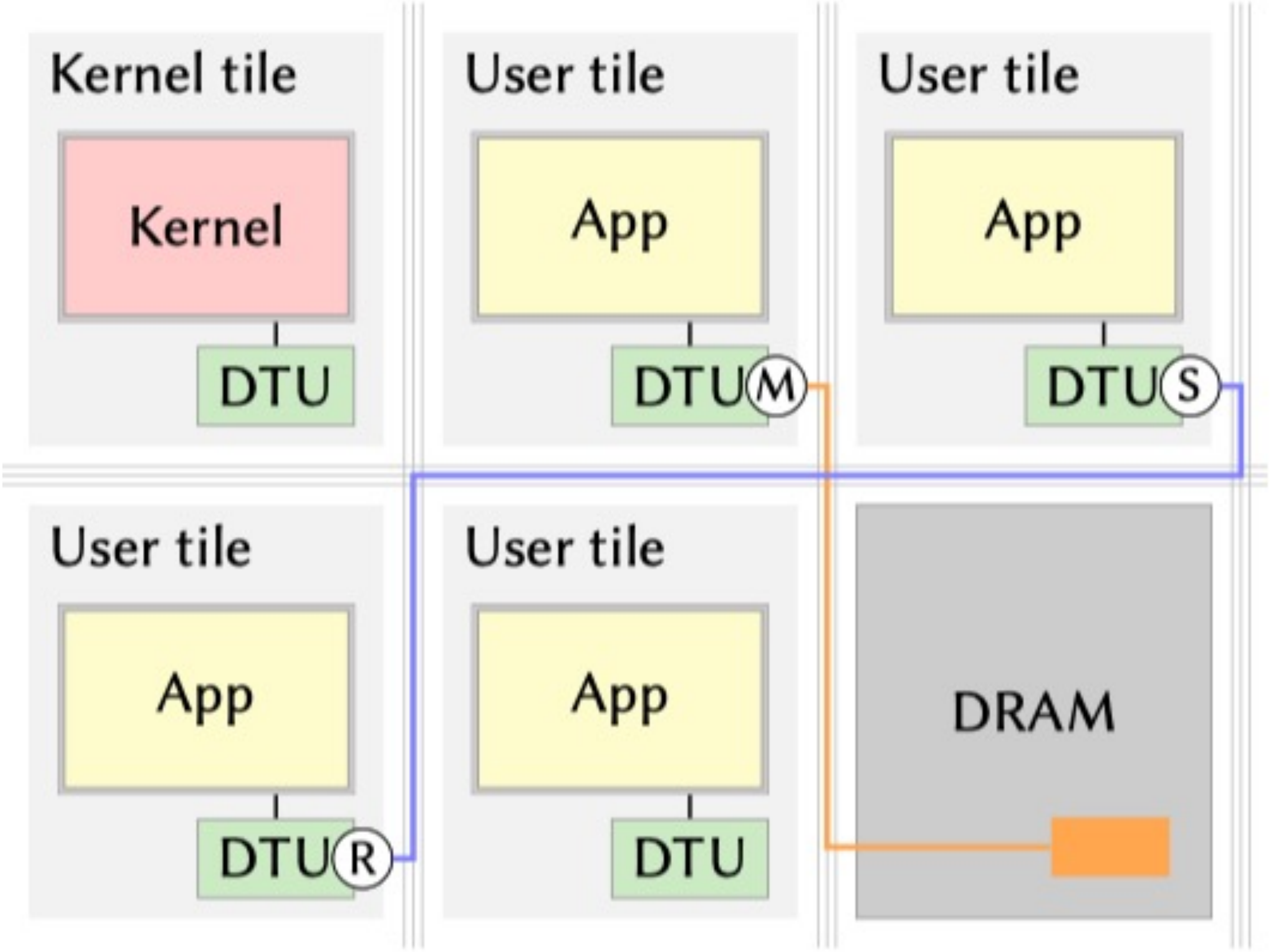


# System Architecture and Background

- Key Ideas of M$^3$[1]:
  - DTU as uniform interface
  - Kernel controls user tiles remotely

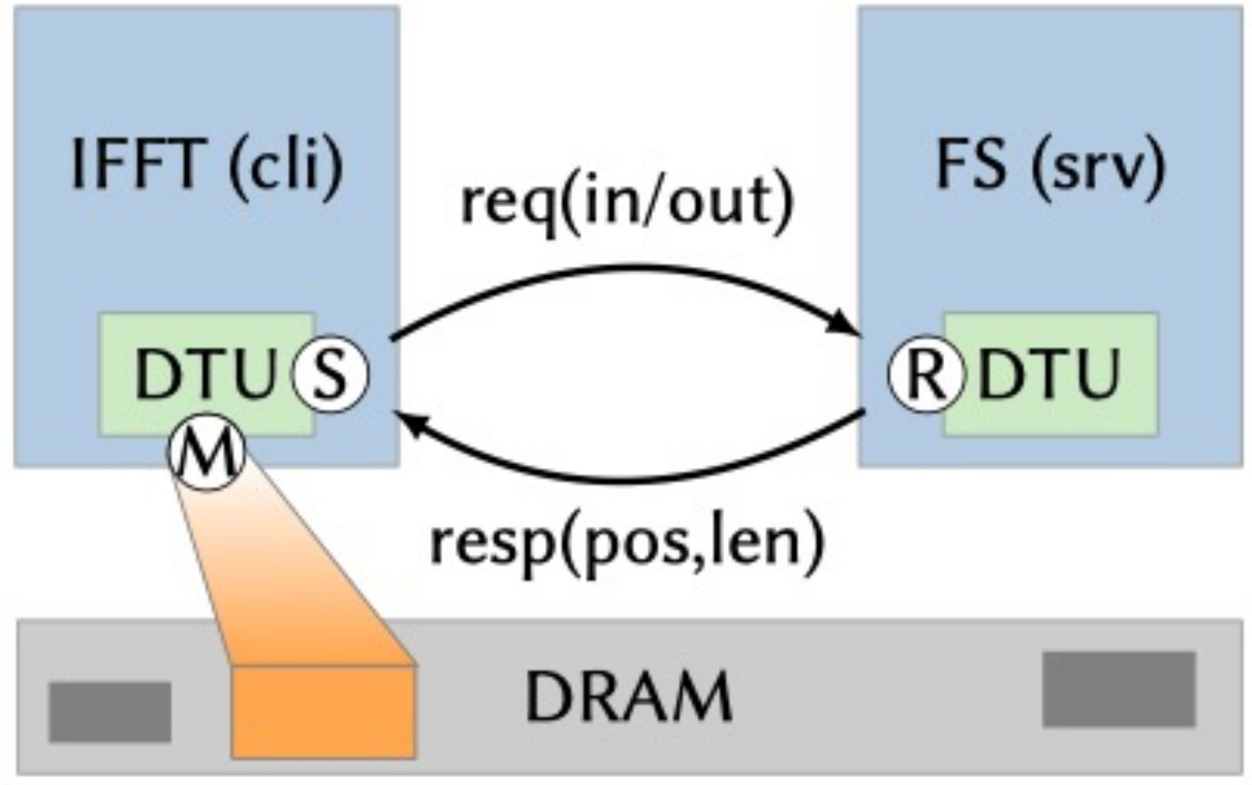
[1]: M 3 : A Hardware/Operating-System Co-Design to Tame Heterogeneous Manycores, ASPLOS 2016

# System Architecture and Background

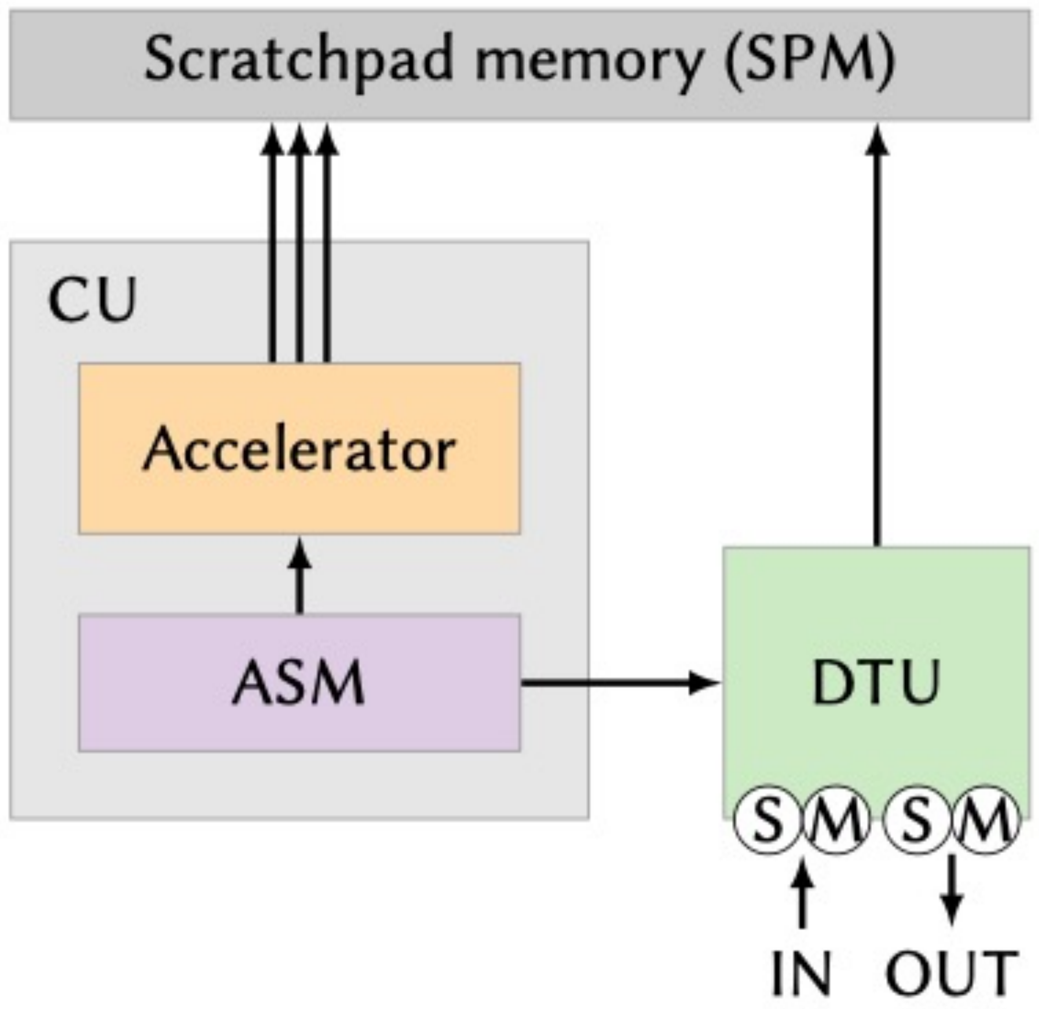


- DTU provides endpoint to:
  - Access memory (contiguous range, byte granular)
  - Receive messages into a receive buffer
  - Send messages to a receiving endpoint

# Autonomous Stream-Processing Accelerators



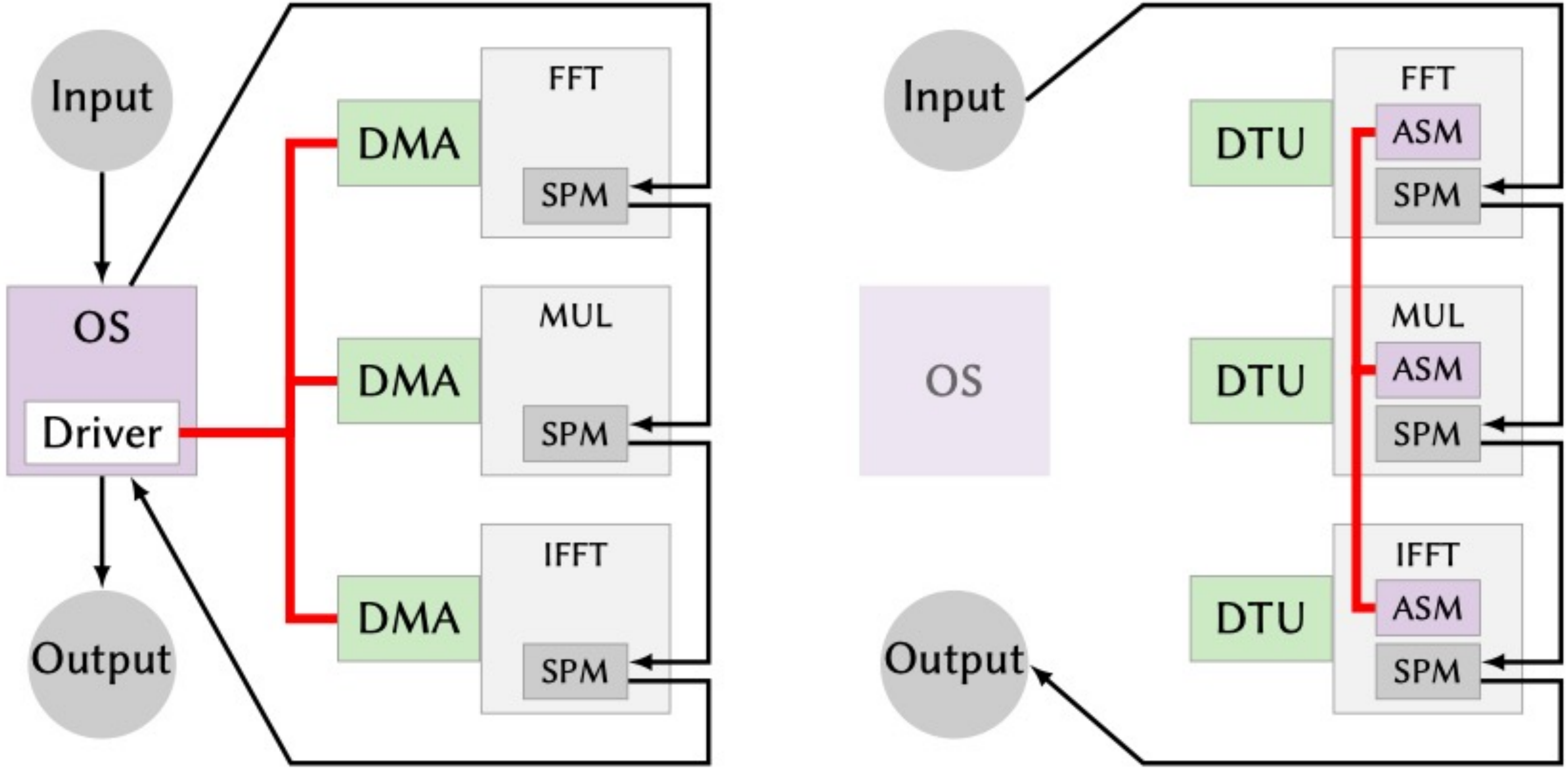
- File protocol
  - Data in memory
  - Msg channel between client and server
    - Req (in) for next input piece
    - Req (out) for next output piece
  - Server configures client’s memory EP
  - Client accesses data via DTU
  - Used by all CUs



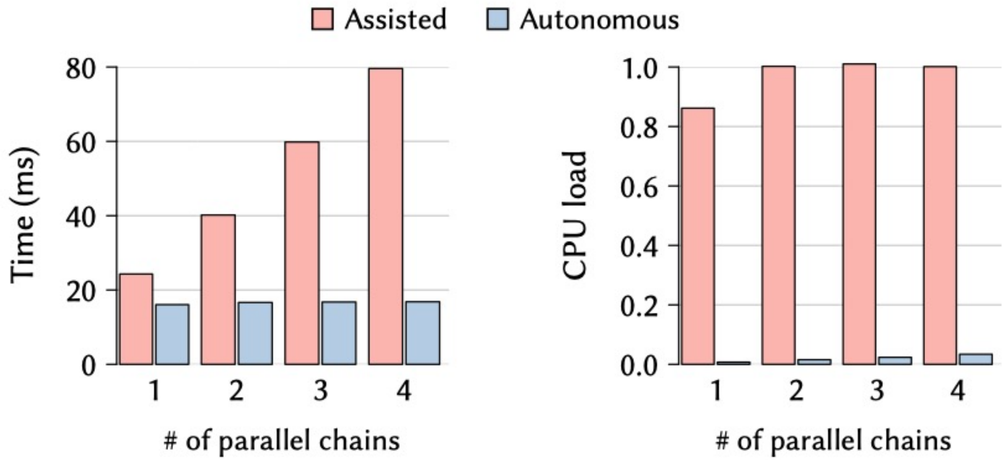
# Autonomous Stream-Processing Accelerators

- Off-the-shelf accelerators
- Accelerator Support Module (ASM):
  - Interacts with DTU and accelerator
  - Implements file protocol for input and output channel

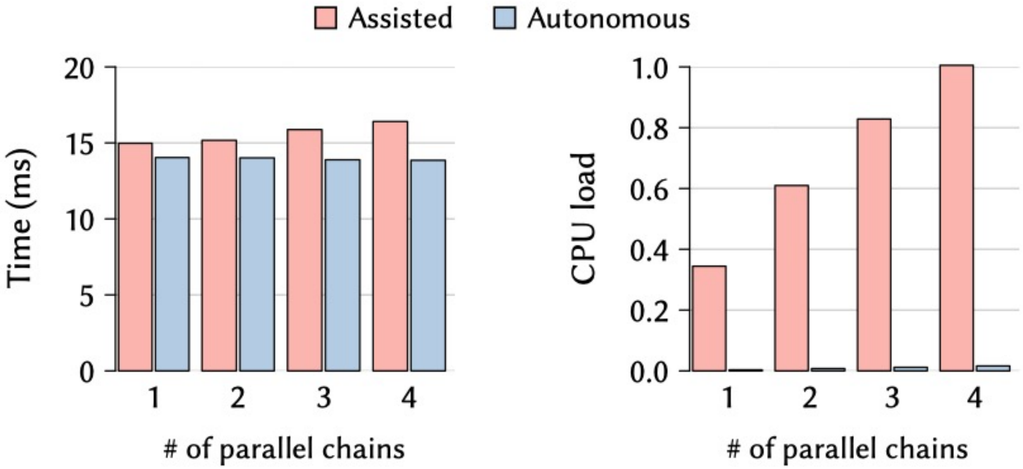
# Assisted vs. Autonomous

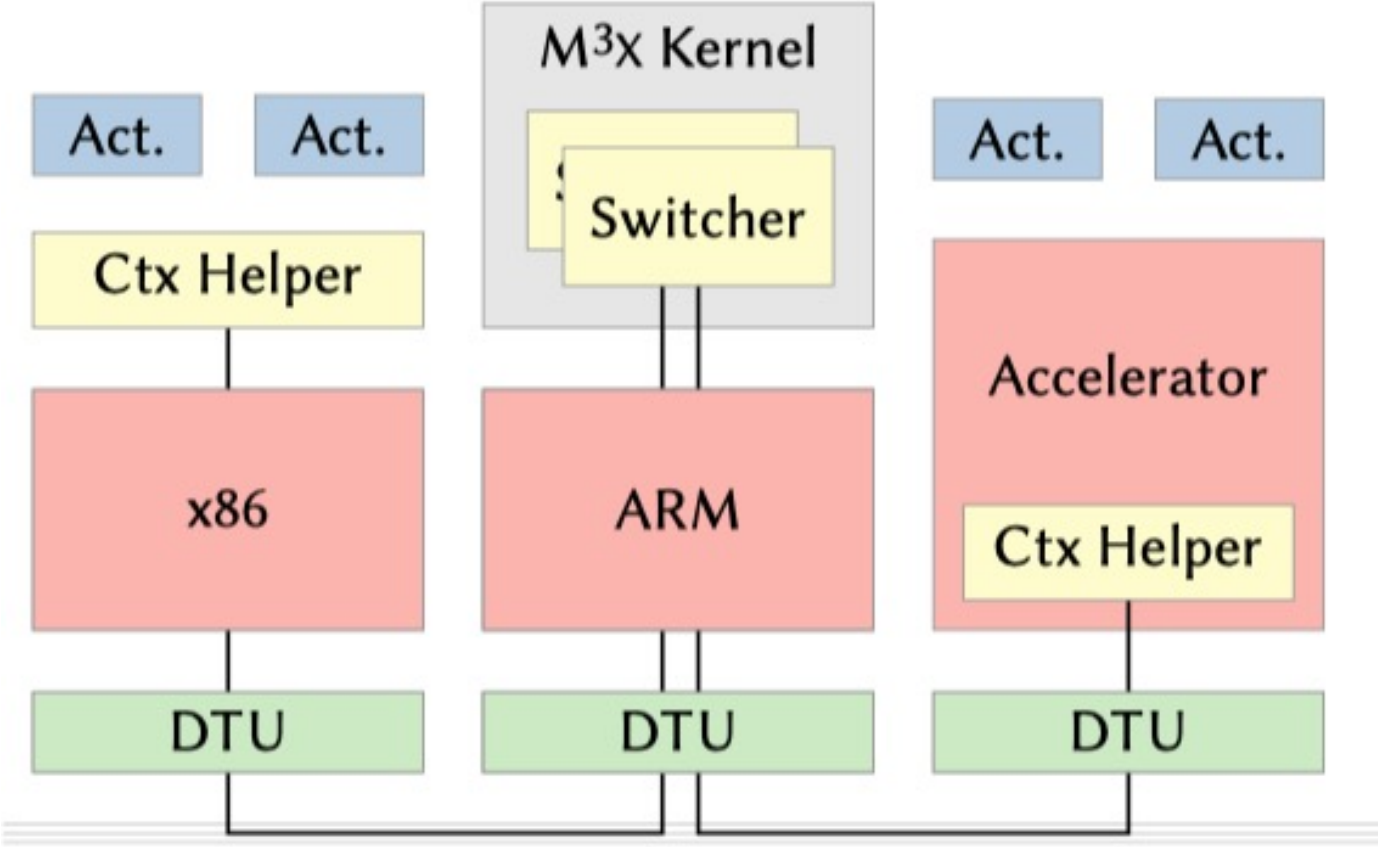


## Accelerator Chains: Results (PCIe-like Latency)



## Accelerator Chains: Results

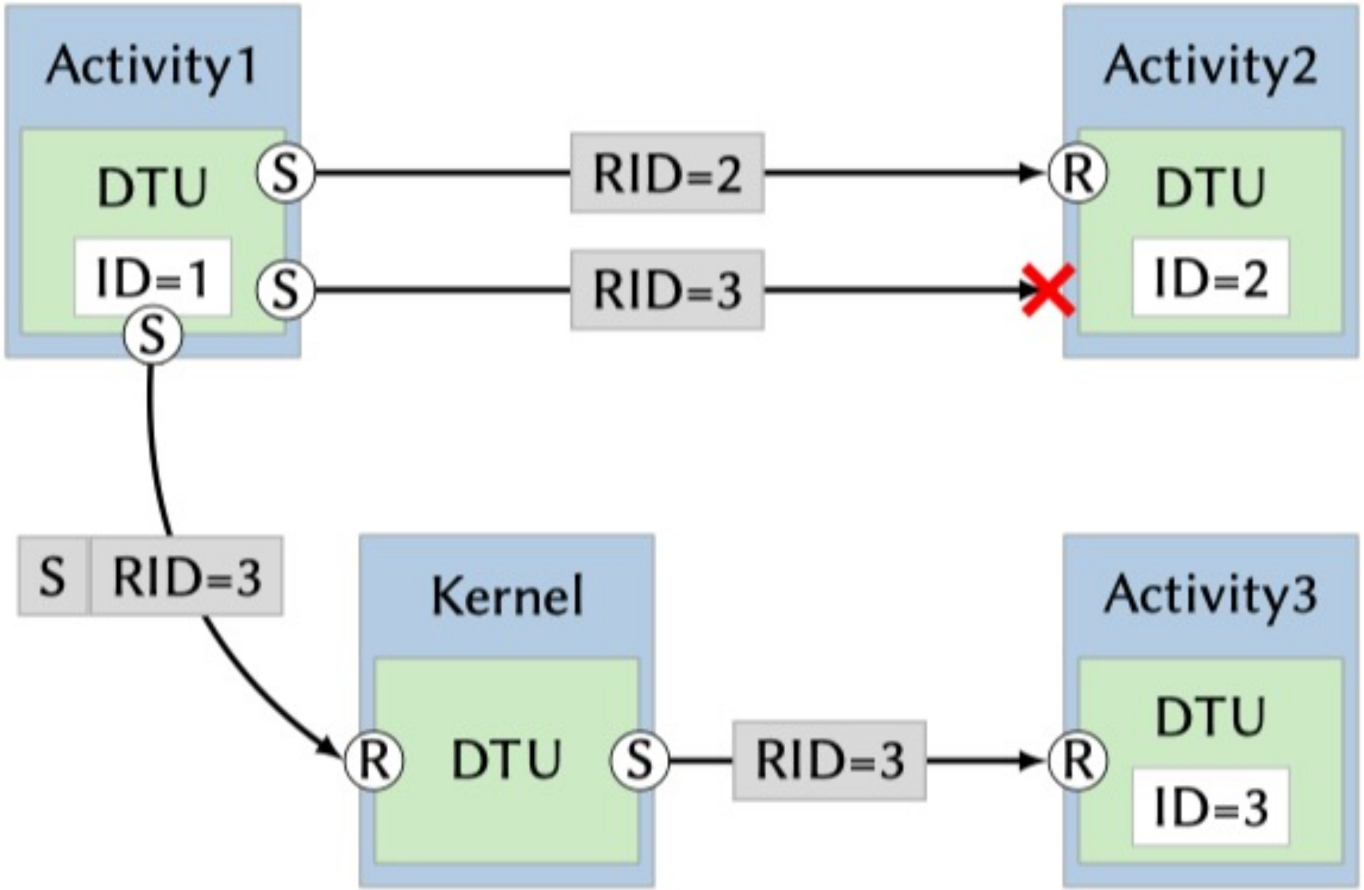




# Fast-Path Communication vs. Context Switching

- Kernel handles complex part
    - Schedules/migrates activities
    - Initiates context switches
- Helper on user tiles implements save/restore
    - General purpose tiles: Software helper
    - Accelerator tiles: Helper implemented in hardware as part of ASM

# Combining Fast-Path Communication with Context Switching



# Conclusion

- M 3 X enables autonomous accelerators and combines fast-path communication with context switching:
  - Adding uniform interface to compute units
  - Using simple and generic protocols
  - Adding lightweight component to accelerators
- Reduces CPU load by a factor of 30
- Retains advantages of fast-path communication
- Uses hardware resource efficiently