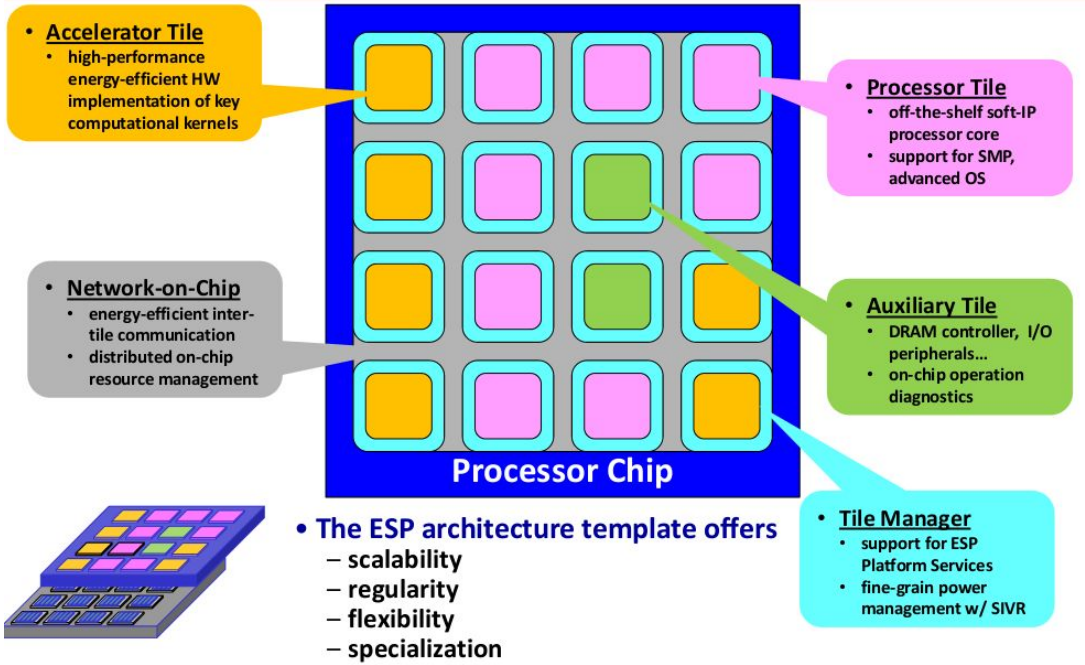
Glue: A Scalable Memory Hierarchy with HLS

Chae Jubb

ESP Architecture

- Heterogenous Architecture
- Multiple CPUs
- Hardware Accelerators
- Memory Controller
- Not Coherent!

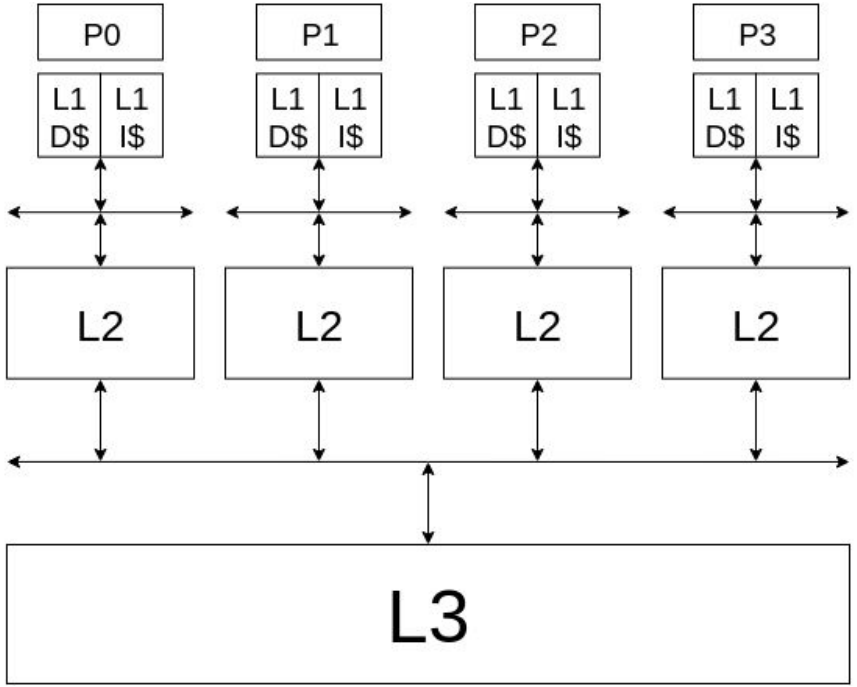


Directory-Based MESI Protocol

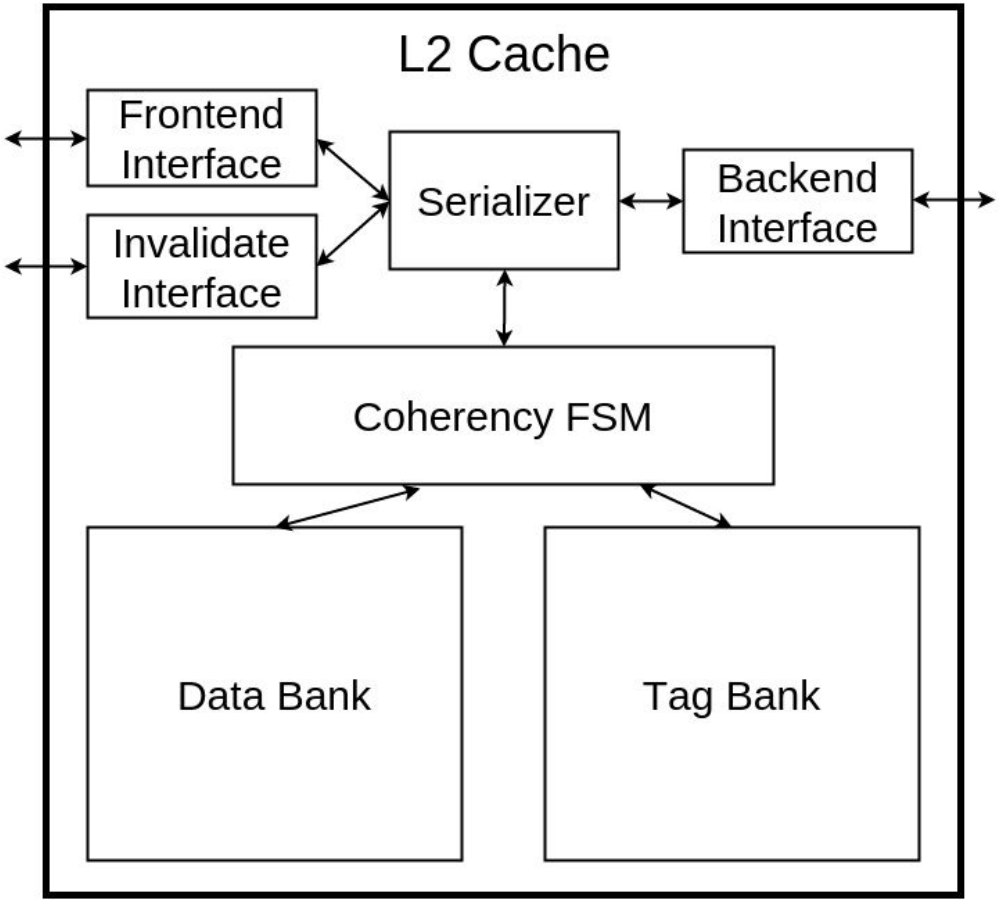
- Four Primary States
  - Modified
  - Exclusive
  - Shared
  - Invalid
- Directory-Based
  - Higher Latency
  - Scalable

Planned Memory Hierarchy

- Additional Level 2
  - Private
  - Write-back
- Additional Level 3
  - Shared
  - Write-back



L2 Architecture

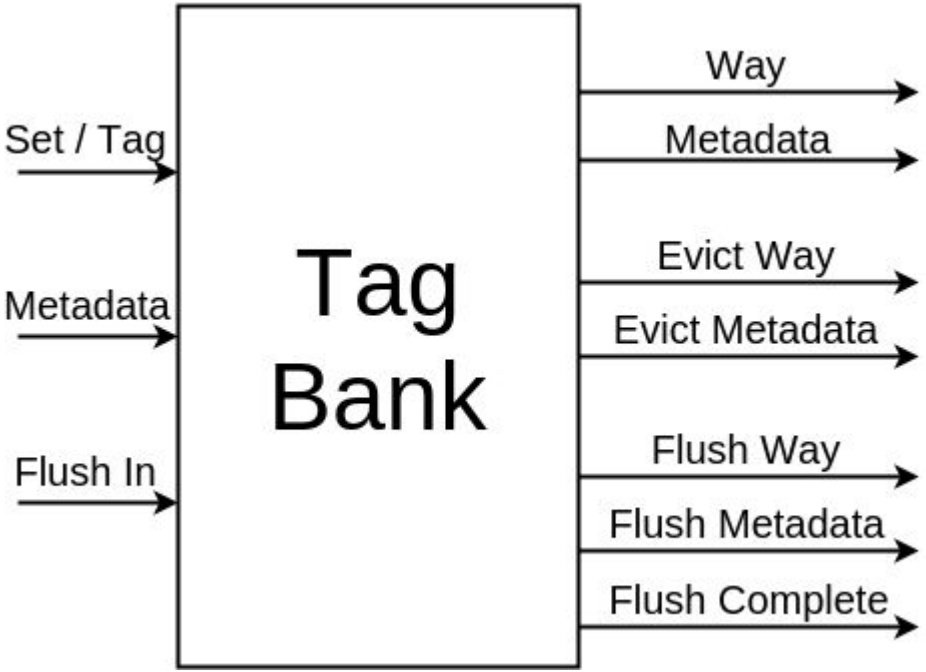


L2 Coherence FSM

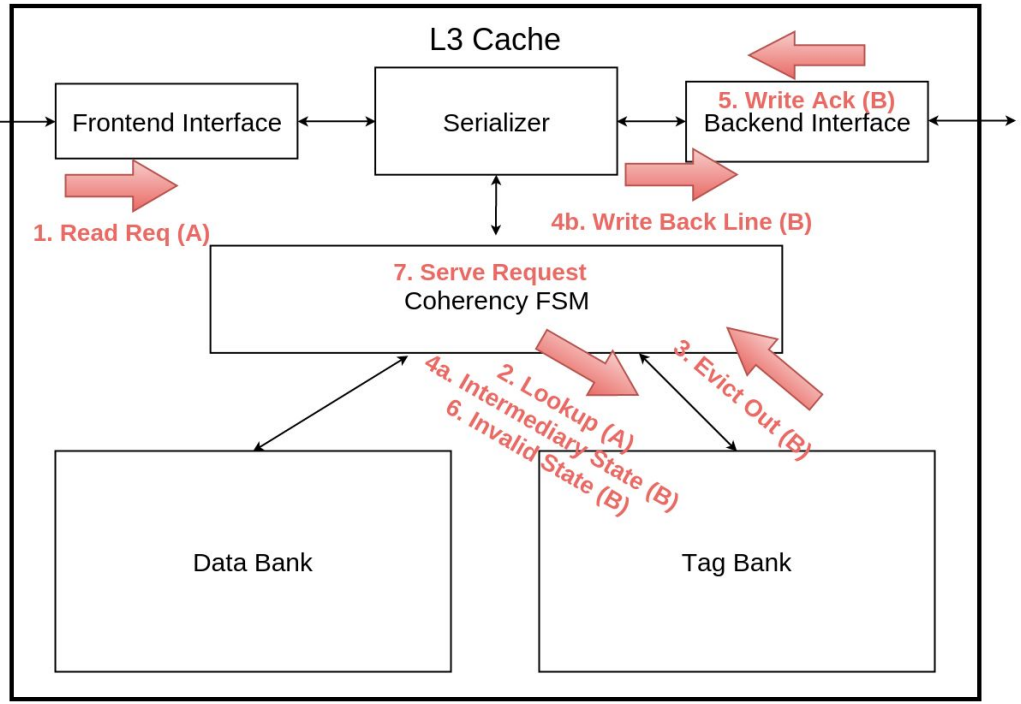
- Responsibilities
- Control Flow
- Single-Core Optimizations
- Forward Plane Stalling
- Read and Write Buffers
- Flushing

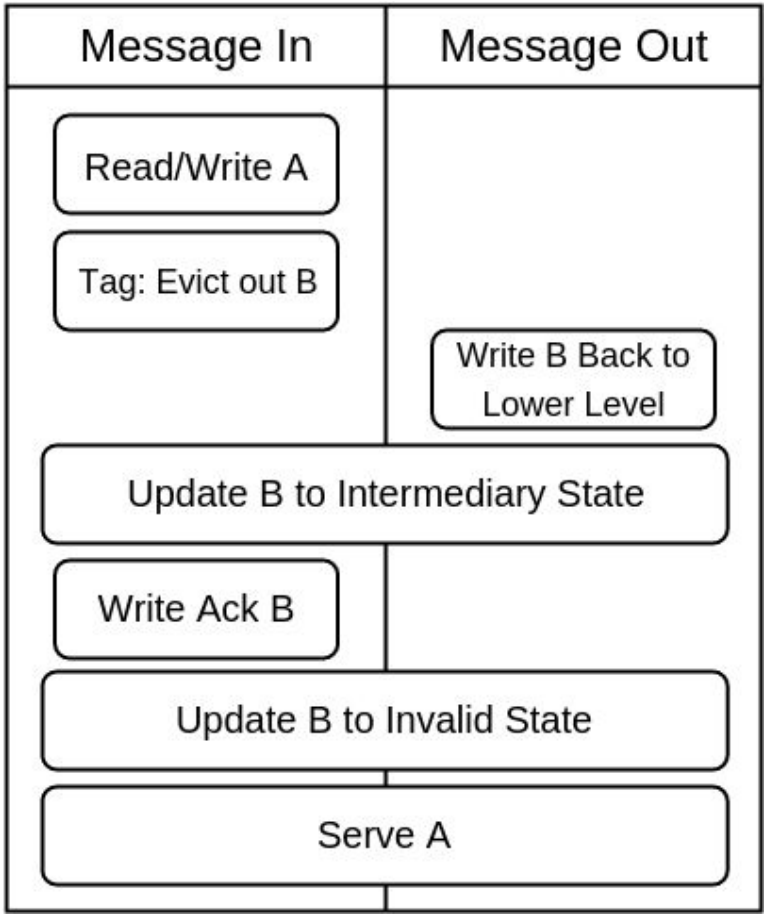
L2 Tag Bank

- Three Planes
  - Nominal
  - Eviction
  - Flushing
- Metadata
  - Tag Bits
  - Coherence State
  - etc



L2 Eviction


L3 Cache
Frontend Interface
Serializer
5. Write Ack (B)
Backend Interface
1. Read Req (A)
4b. Write Back Line (B)
7. Serve Request
Coherency FSM
2. Lookup (A)
4a. Intermediary State (B)
6. Invalid State (B)
3. Evict Out (B)
Data Bank
Tag Bank


Message In
Message Out
Read/Write A
Tag: Evict out B
Write B Back to Lower Level
Update B to Intermediary State
Write Ack B
Update B to Invalid State
Serve A

Flushing the L2

- Entire Cache
- Three Phases
  - Initiation
  - Flushing
  - Completion
- Led by tag bank

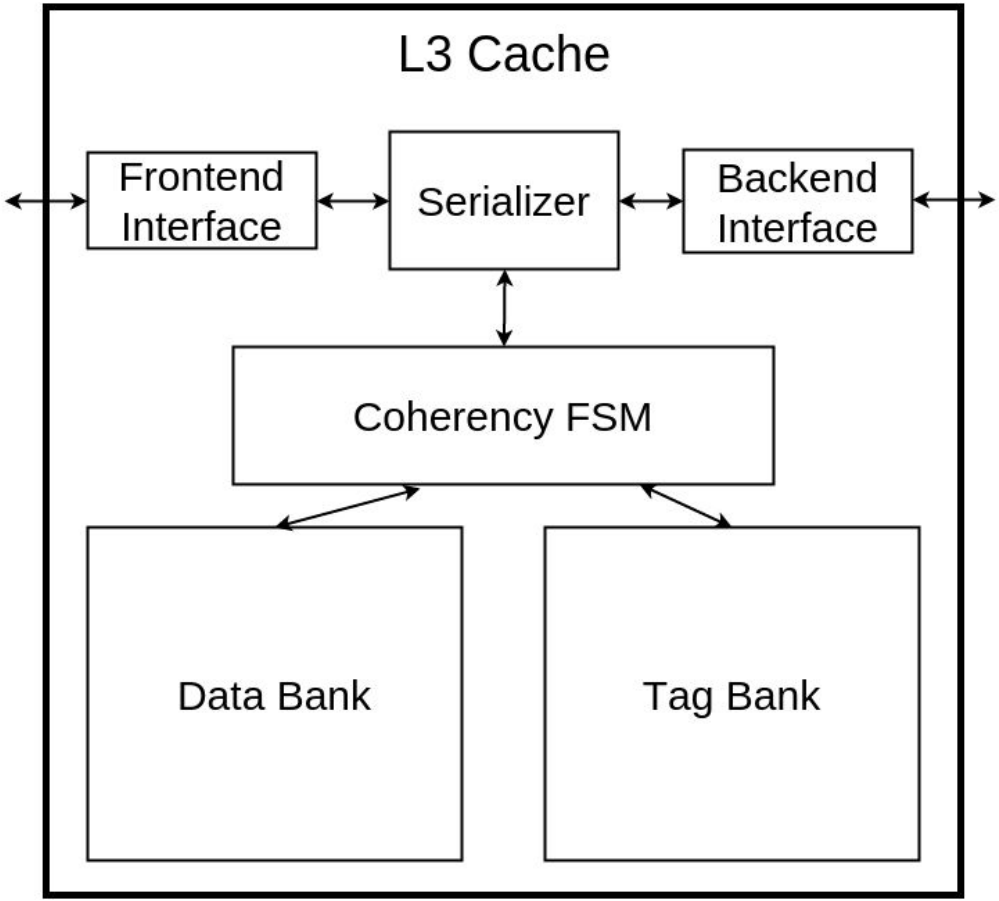
- Similar flow as eviction
- Flushing occurs in background
  - Tag Bank
  - Coherence FSM

L2 Data Bank

L2 Invalidation

- Flushing from processor direction
- Invalidation from main memory direction
- Separate plane

L3 Architecture



L3 Coherence FSM

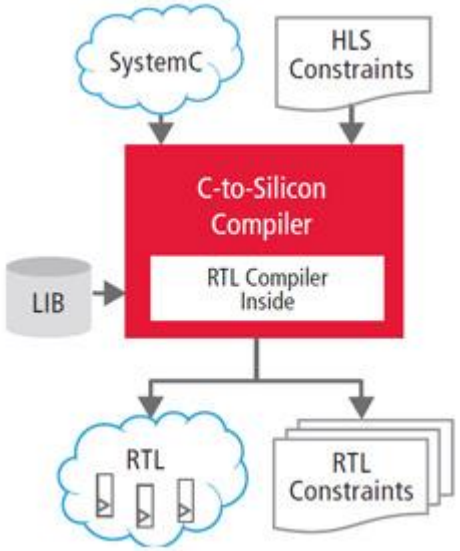
- Serves as directory
- Modify given protocol
- Series of non-blocking actions
- Intermediary States
  - Only Reads
  - Only Writes
  - Reads and Writes
- Dirty Bit
  - Unnecessary in L2

Validation Plan

- Toolbox
  - HLS Specification of DUT
  - HLS Specification of test bench
  - RTL Implementation of DUT
- Validation and Verification
  - HLS DUT / HLS TB
  - RTL DUT / HLS TB
  - Formal Verification of RTL DUT
- Coverage
  - L2 Tag Bank
    - 100% Line
    - 88% Branch
  - Entire L2
    - 94% Line
    - 63% Branch
- L2 Tag Bank Latency Upper Bounds
  - Arbitrary Request: 6 cycles
  - Read Hit: 3 cycles

High Level Synthesis: Why?

- Easy Parameterization
  - Number of Sets
  - Number of Ways
- Quick Iteration
- Latency Insensitive Interface
- I’m better at it

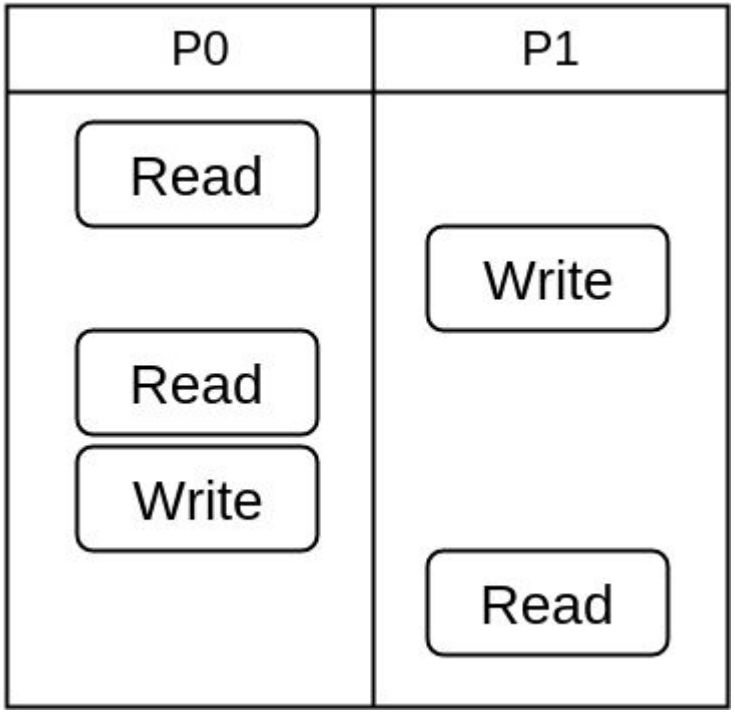


HLS Experiences

- Force a CAM
- Constrain Latency
- “Protocol Region”
- Elimination of data bank
- Location on Pareto curve

Memory Hierarchy Testing

- Multiple L2 caches
- Single shared L3 cache
- Create cache interconnect
- RTL simulation
- IT WORKS!

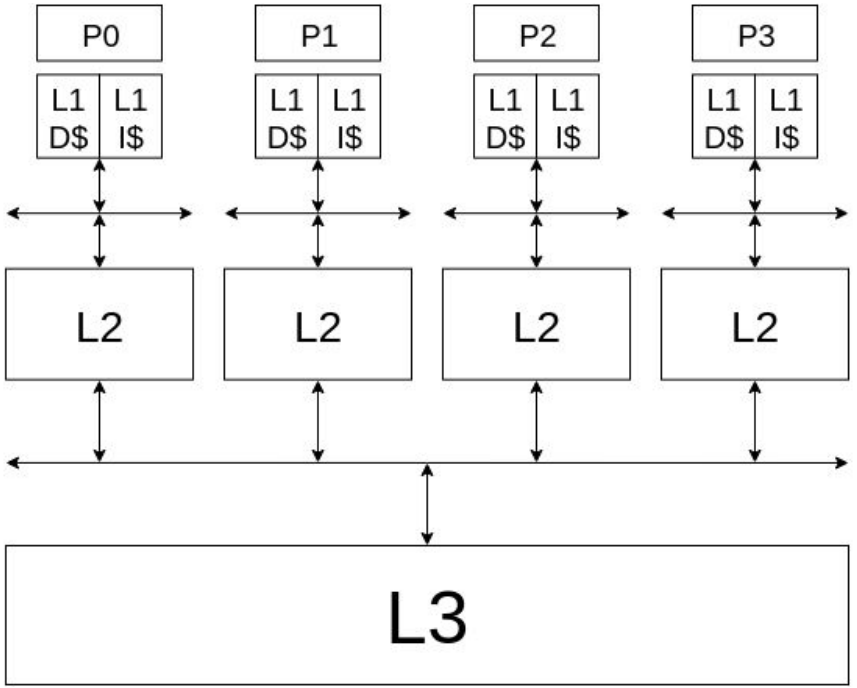


Future Work

- Flushing
- Concurrent Requests to Valid Line (L3)
- Forward Plane Stalling (L2)
- Miscellaneous race conditions

Summary

- Coherent Memory Hierarchy
  - MESI Protocol
  - Multiple Private L2s
  - Single Shared L3
- Validation and Verification
  - High coverage
  - Formal latency bounds on tag bank
- High Level Synthesis
  - Aggressively optimized for speed
  - Parametrization



Demo of Full Simulation!