# Token Coherence: Decoupling Performance and Correctness

Milo M.K. Martin, Mark D. Hill, David A. Wood
University of Wisconsin-Madison
Presented for ECE259 by Bryan Fleming

# Motivation

- Ahmdal’s Law – make the common case fast and the uncommon case work
- Cache-to-cache misses are the common case
- Directory protocols require indirection
- Snooping protocols require ordered network

# Token Coherence Overview

- Targeted for medium-sized systems
- Avoid indirection latency and not require an ordered network
- Correctness substrate based on tokens ensures all cases work
  - Ensure safety
  - Avoid starvation
- Performance protocol makes common cases fast
  - Makes requests to substrate

# Token Coherence Implementation

- Token invariants ensure correctness
- MOESI state deterministic based on the number and types of tokens held
    - No need for complicated state machine for transients
- Persistent Requests avoid starvation

- Can now exploit fast, unordered network

# Evaluation

- Simulated against Directory and Snooping
- Really low reissues and persistent requests!
- Less bandwidth-limited than snooping due to different network

# Discussion Questions

- Is Token Coherence really faster and simpler than directories and snooping?
- Can we afford the bandwidth overhead?
- Are these workloads representative?
  - How would ocean perform?

- How might we apply decoupling idea to larger systems?