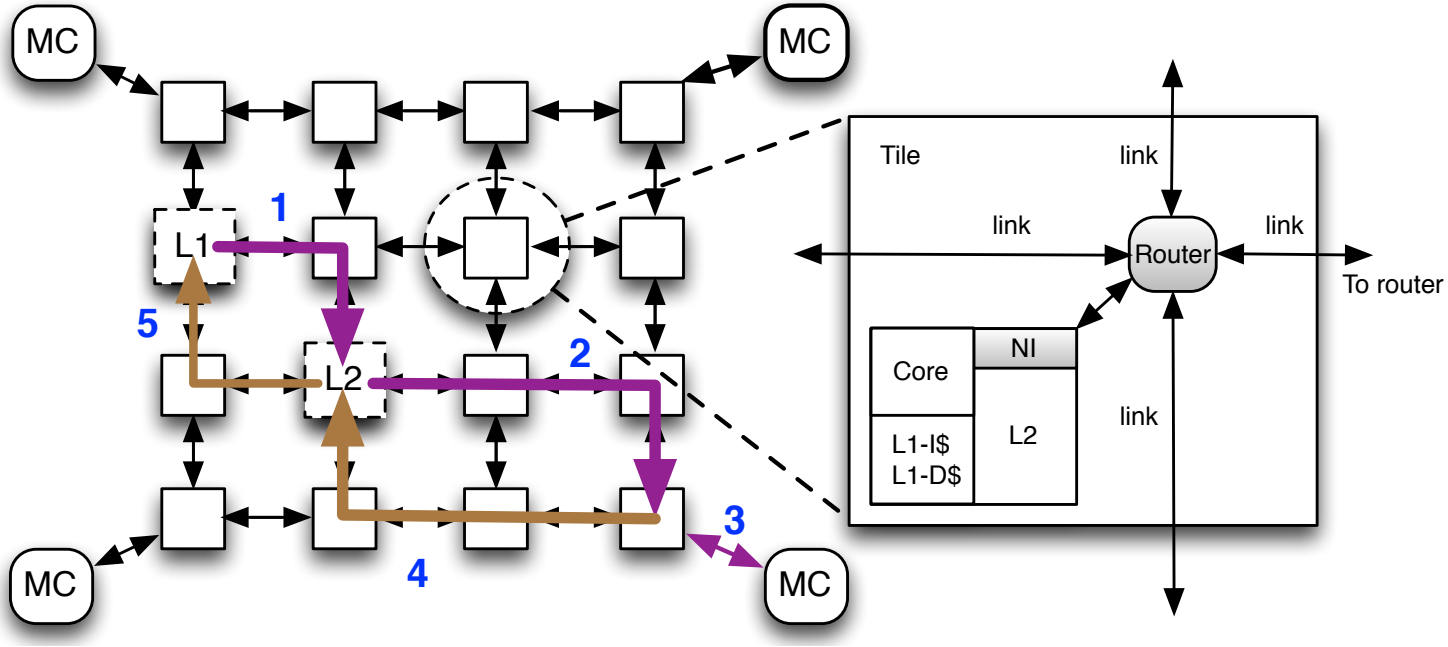
# NoΔ: Leveraging Delta Compression for End-to-End Memory Access for NoC-Based Multicores

**Jia Zhan**[1], Matt Poremba[1], Yi Xu[2], and Yuan Xie[1,2]

**[1]The Pennsylvania State University**
**[2]AMD Research China Lab**


# Network-on-Chip (NoC) Based Multicore




1: L1→L2 due to L1 miss

2: L2->MC due to L2 miss

3: Off-chip memory access

4: MC→L2 for L2 updates

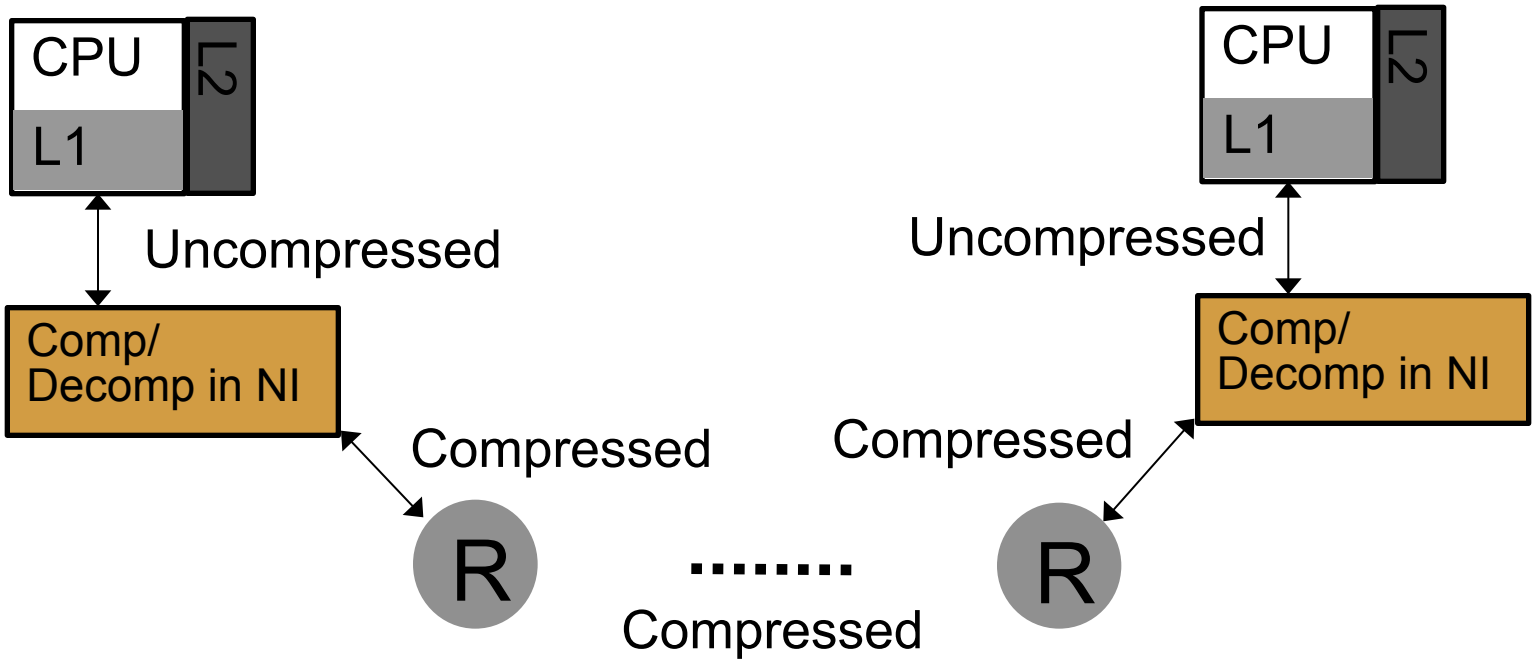
5: L2→L1 for access to the core

# NoC Optimization Techniques

- Challenges:
  - Long end-to-end latency
  - High network power
- Conventional optimization techniques:
  - Router microarchitecture
  - Network topology
  - Routing algorithm, etc.
- An orthogonal & complementary approach:
  - **Data Compression:** compresses data messages into more compact bits before they are stored or transmitted.
  - Reduces network traffic, benefits both latency & power.

# Data Compression

- Two kinds of hardware compression techniques:
  - Cache Compression
    - expands the cache capacity.
    - requires modifications to the cache architecture, however.
  - NoC Compression
    - Compress/Decompress packets at the Network Interface (NI).

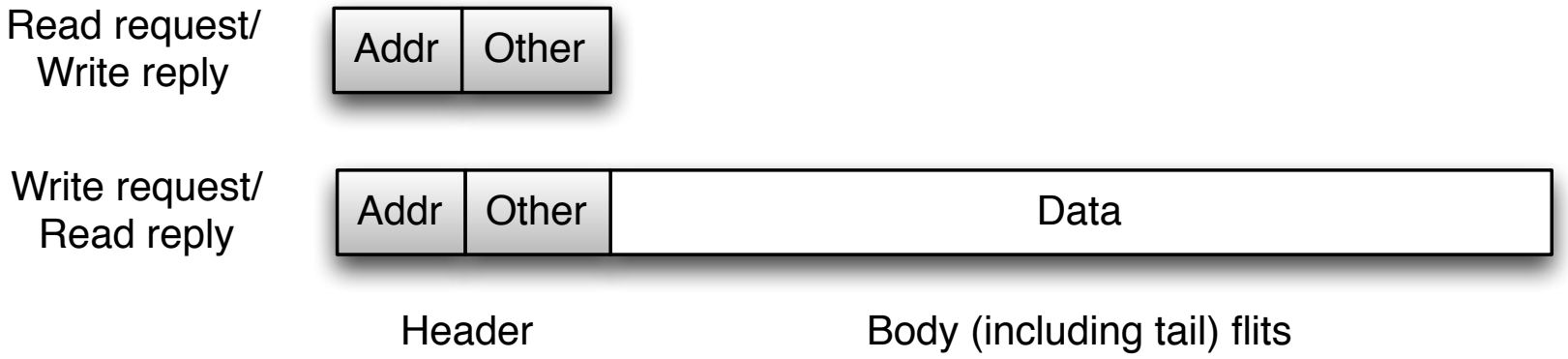


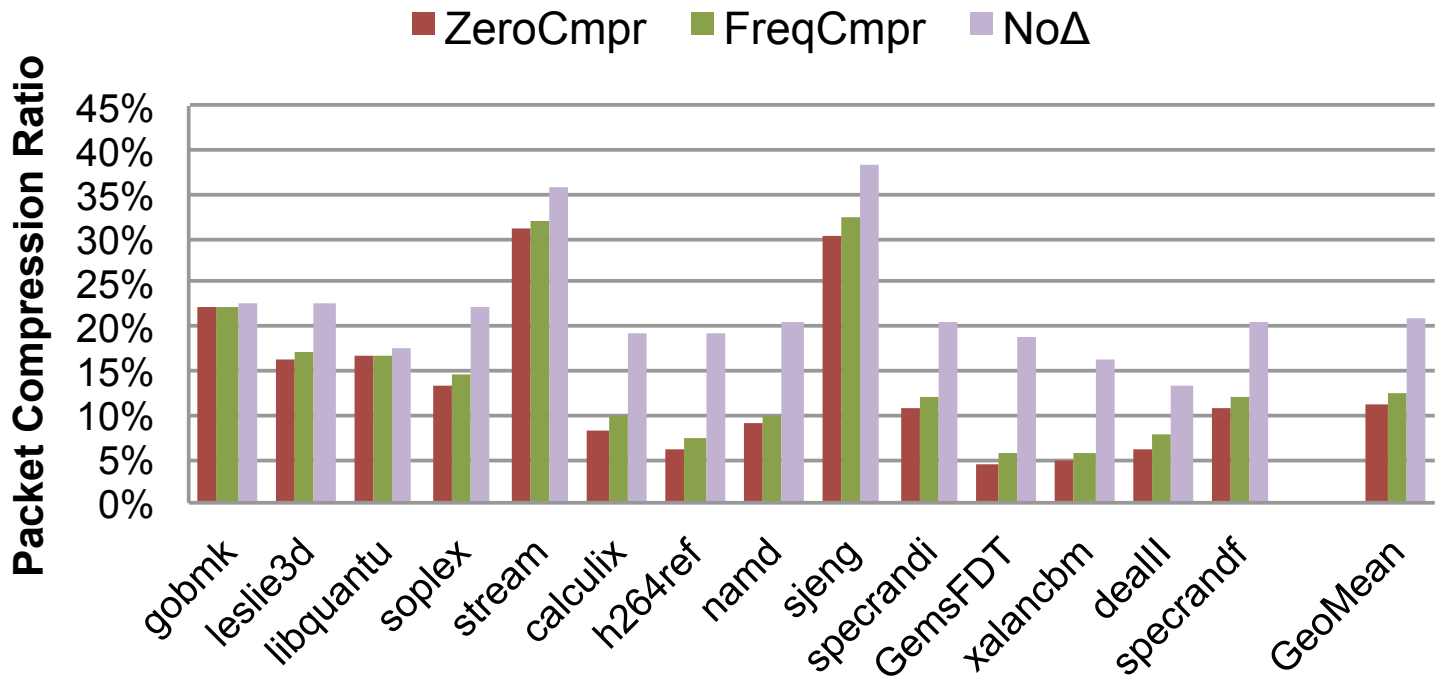
# NoC Compression

- Existing approaches:
  - Zero-based compression (ZeroCmpr)
  - Frequent value compression (FreqCmpr)

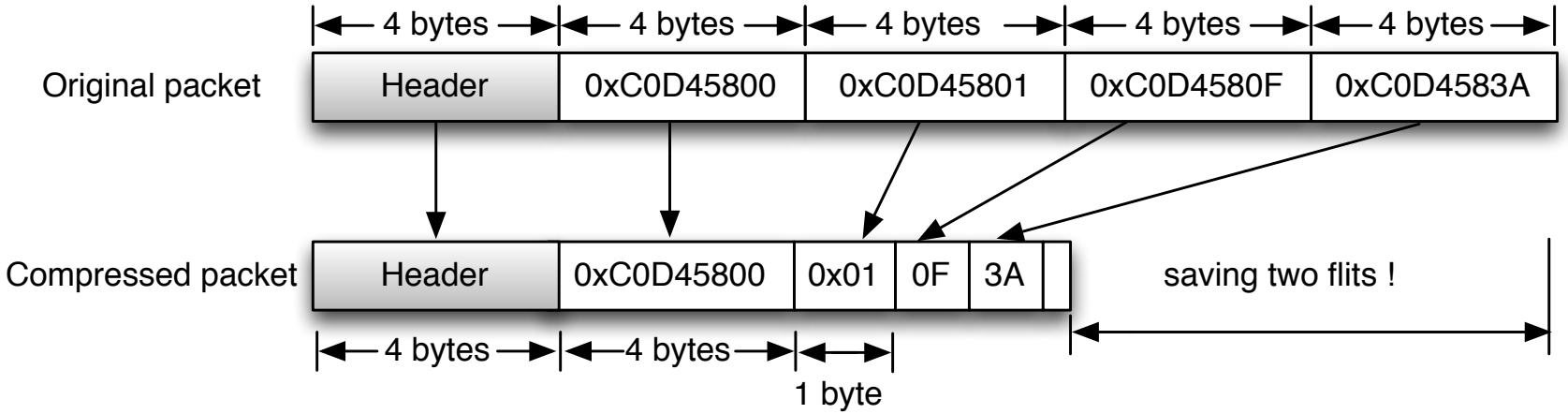
- Our approach: **Delta Compression**
  - transmits data in the form of differences rather than the complete data set
  - Widely applied in video compression and Internet traffic reduction.
  - Can be applied in NoC:
    - Network Interface
    - Compression/Decompression modules

# Packet Compression



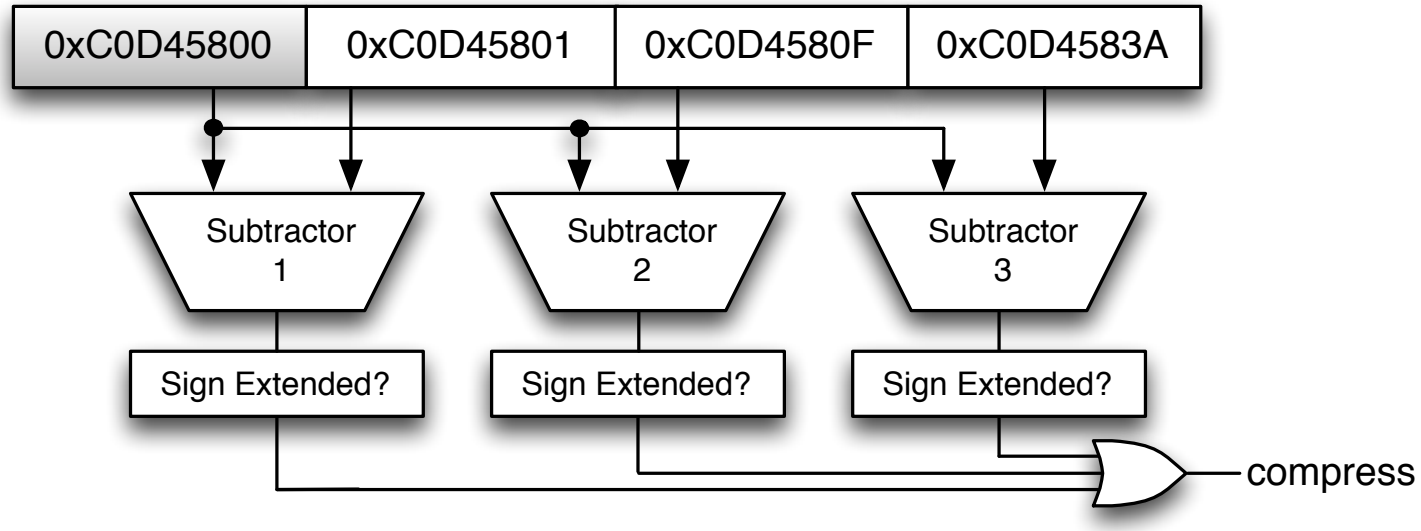
# An example




- The example fixes the base length to one flit!
- Alternatively…
  - Variable-length base
  - Multiple bases

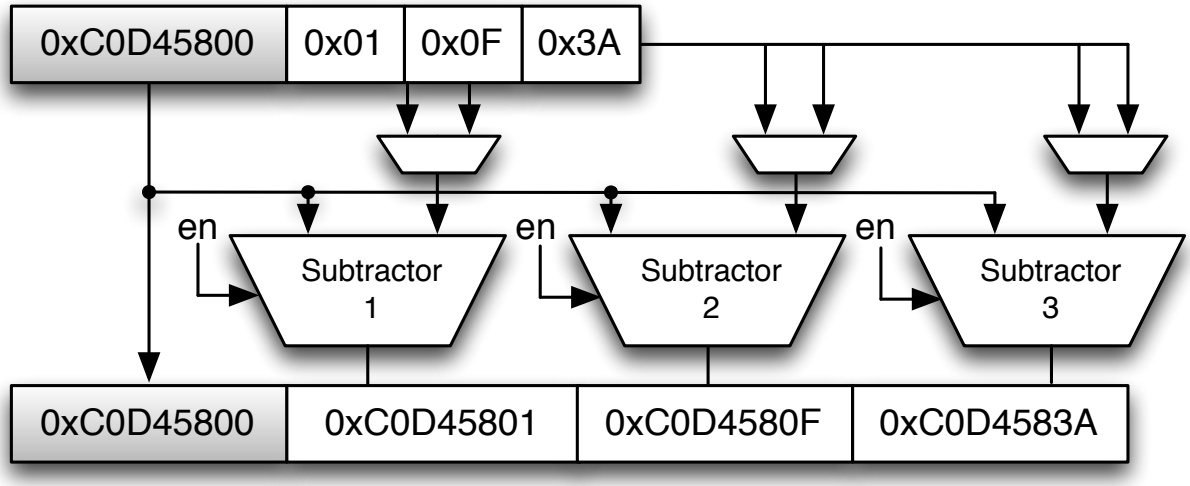
# Architecture Support

- Compression module



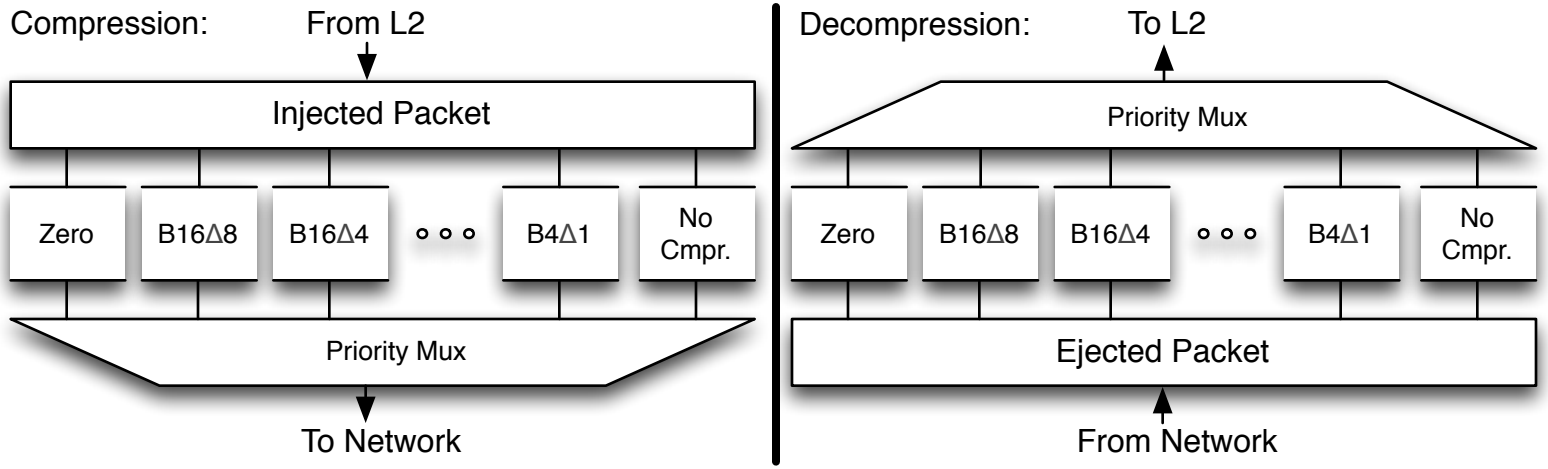
# Architecture Support

- Decompression Module



# Architecture Support

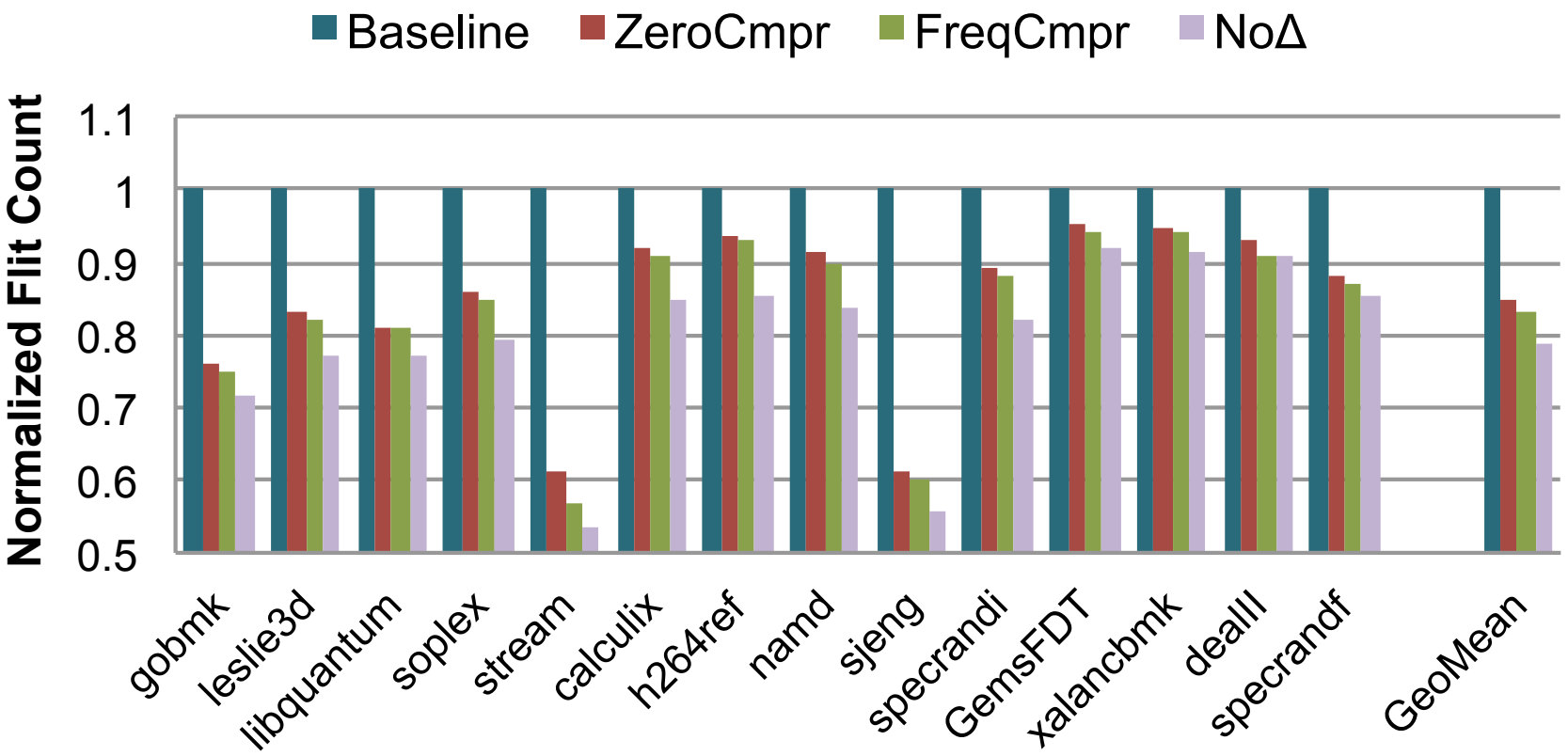
- NoΔ Architecture



# Experimental setup

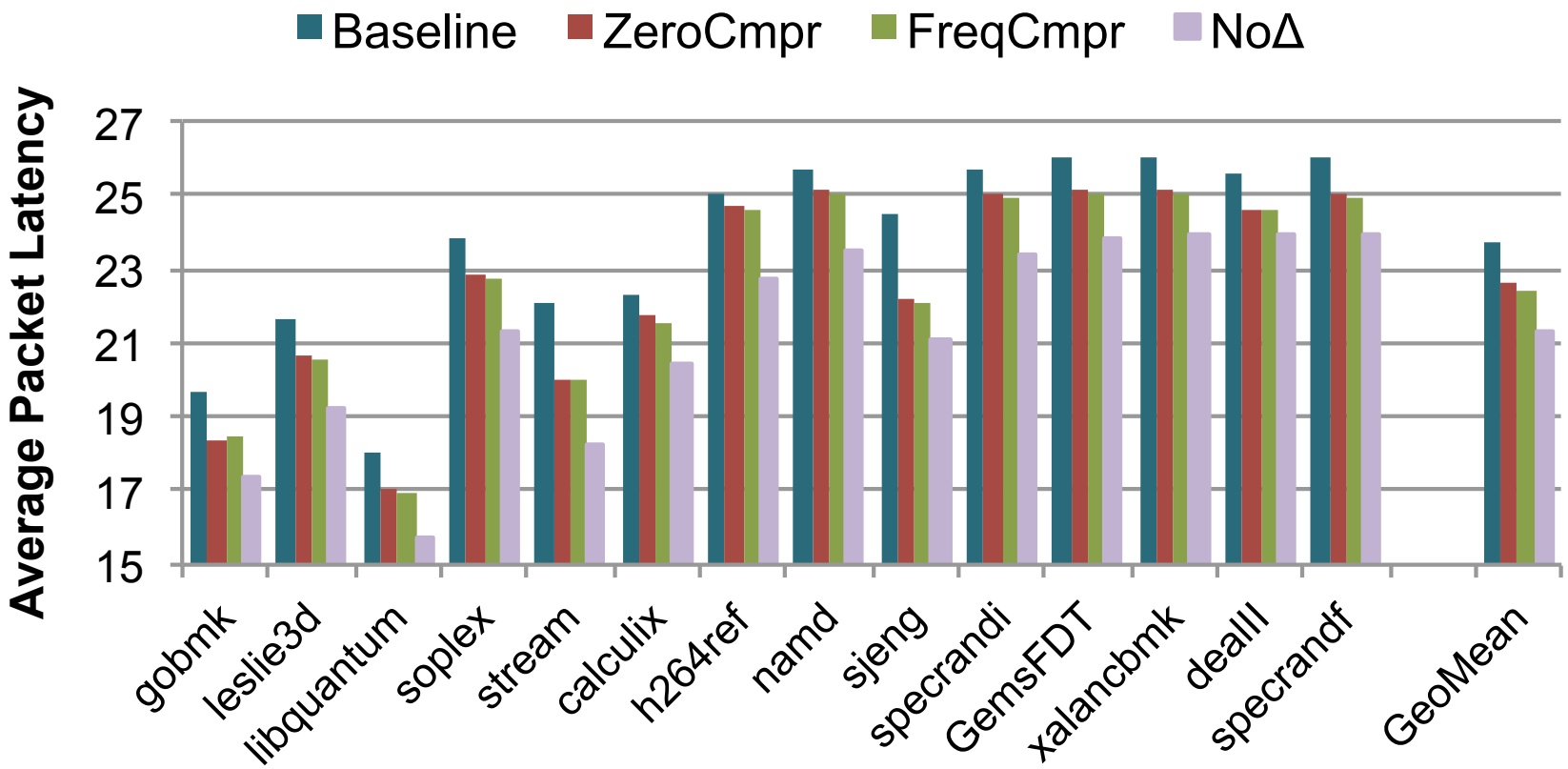
- Gem5 full system simulator
- Ruby: detailed memory system simulation
- Garnet: network simulation
- DSENT: network power analysis
- Detailed configuration:
  - 16 cores, 2GHz
  - 8KB 2-way L1 I&D cache, 16x1MB 8-way L2 cache
  - 4GB memory, 4 memory controllers
  - Cache protocol: MESI
  - Network: 4x4 mesh, DOR routing
  - Classic five-stage router, 2 VCs per port, 4-flit depth per VC
  - 5 flits per packet, 16 bytes per flit

# Network Traffic Reduction



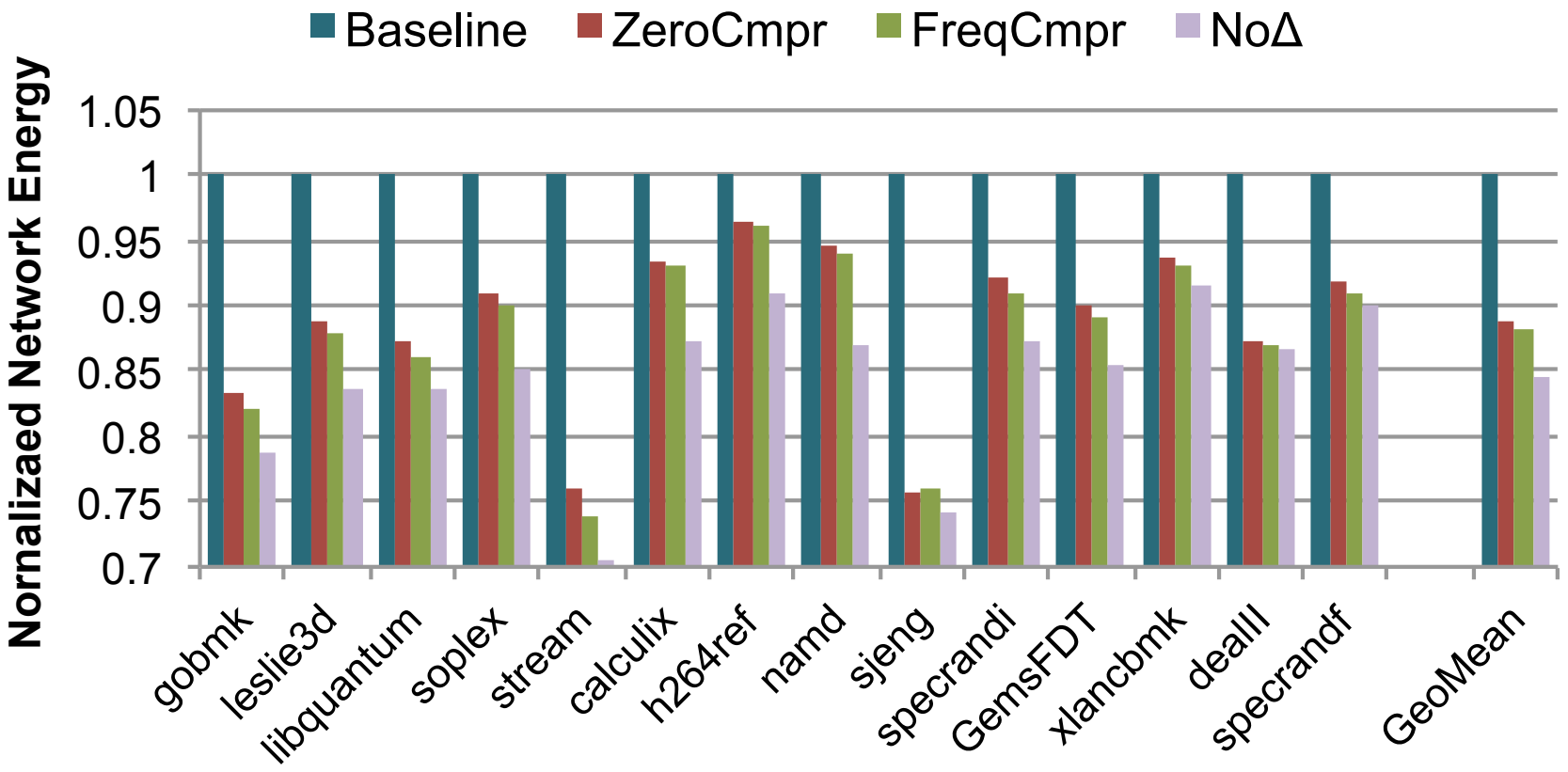

Flit reduction by 21.1% on average.

# Packet Latency




Packet latency cut by 13.1% on average.

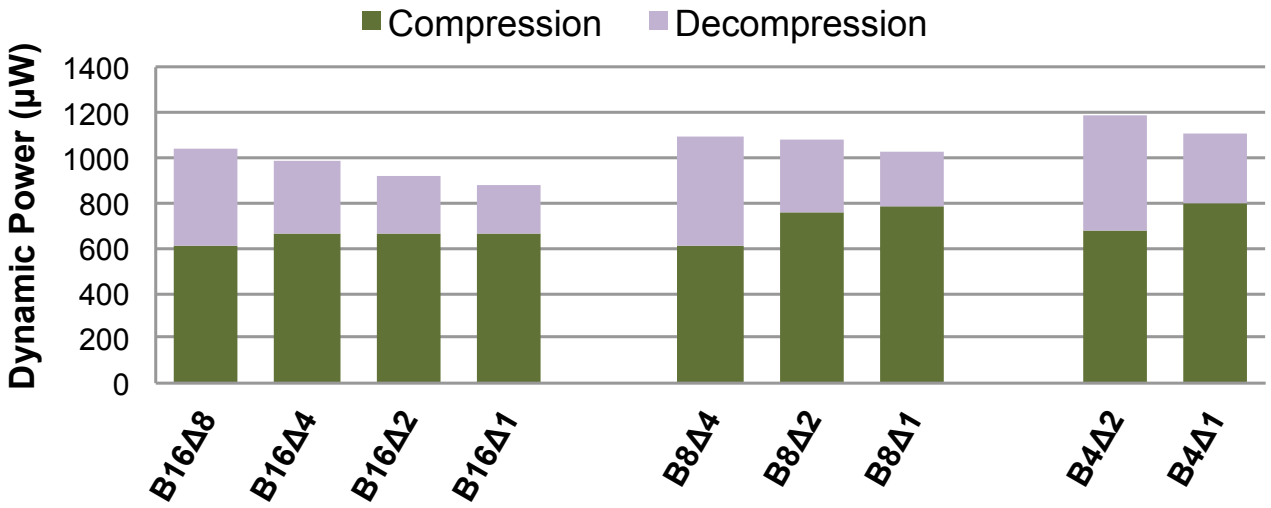
# Energy savings




Average network saving: 15.3%

# Overhead estimation

- Power of compression/decompression modules
  - 0.69% - 1.81% overhead of a router




- Area overhead of compression/decompression modules
  - 0.006mm$^2$
  - Less than zero compression, slightly larger than the best case of frequent value compression

# Summary

- NoCs face challenges of both latency & power.
  - Data compression can be integrated in the NI to compress network traffic to mitigate these challenges
- Existing approaches use zero compression or frequent value compression
- We propose NoΔ that transmits packets in the form of differences between sequential values.
- Experimental results shows 21.1% flit reduction, which leads to 13.1% latency drop and 15.3% energy saving on average.
- In the paper:
  - Detailed NoΔ mechanism
  - More sensitivity studies.
  - Detailed overhead comparison with other related work.