

Banshee: Bandwidth-Efficient DRAM Caching via Software/Hardware Cooperation

Xiangyao Yu¹, Christopher Hughes², Nadathur Satish², Onur Mutlu³, Srinivas Devadas¹

MIT¹

Intel Labs²

ETH Zürich³

- In-package DRAM has
 - 5X higher bandwidth than off-package DRAM
 - Similar latency as off-package DRAM
 - Limited capacity (up to 16 GB)
- Use in-package DRAM as a cache
- Banshee maximizes both in- and off-package DRAM bandwidth efficiency to deliver high performance

Idea 1

- Track DRAM cache contents using TLBs and page tables
- A novel, lightweight TLB coherence mechanism with low complexity

Idea 2

- Bandwidth-aware frequency-based replacement policy to reduce unnecessary DRAM cache replacement traffic

Banshee improves performance by **15%** and reduces in-package DRAM traffic by **35.8%** over the best-previous latency-optimized DRAM cache design



ISTC
BIG DATA

ETH zürich