TOHOKU UNIVERSITY
Cyberscience Center
Cyberscience Center, Tohoku University

Memory Architecture for Vector Supercomputers

Introduction

Vector Supercomputers

Vector processor

Efficient computation by vector instructions with long vector length

Memory system

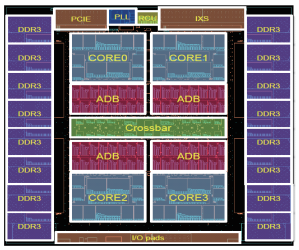
High memory bandwidth by a lot of memory channels

Vector supercomputers can effectively process applications, and are utilized in scientific and engineering fields

Memory System of SX-ACE

High memory bandwidth is crucial to increase the effective performance of applications

- An SX-ACE processor has 16 DDR3 channels
- No more space for increasing memory interfaces in future



Chip layout of SX-ACE Processor

Capacity and bandwidth walls are coming up soon!

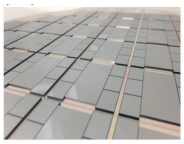
Heterogeneous Memory Architecture

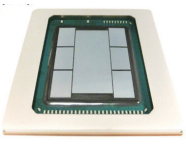
Background

High Bandwidth Memory (HBM) has attracted attentions as a memory module for supercomputers

- Significant improvement of memory bandwidth
- Limited memory capacity due to integration technologies

(E.g. The latest vector supercomputer Succeeder DYSBASIA integrates six HBM modules on a silicon interposer



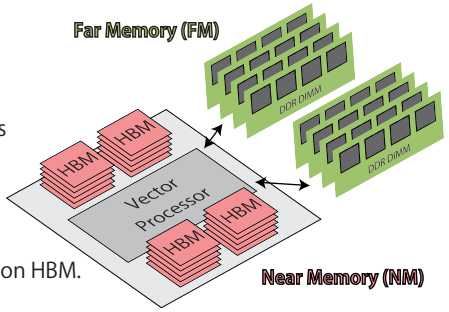


- Bandwidth 1.2TB/s (Larger than SX-ACE)
- Capacity 48GB (Smaller than SX-ACE)

Heterogeneous Memory Architecture (HMA)

Combining different memory modules as a single memory system

- Taking advantages of both modules
  High memory bandwidth by HBM modules
  Large memory capacity by DDR modules
- Challenges



Data Management

Frequently accessed data should be placed on HBM.

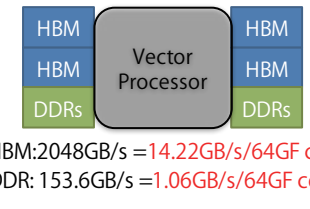
Memory Module Configurations

The number of memory interfaces integrated on a chip is limited. Due to this limitation, the trade-off between memory bandwidth and capacity should be considered.

Bandwidth and Capacity Trade-off

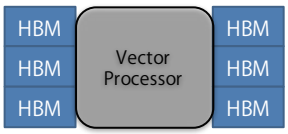
There is a trade-off between capacity and bandwidth. As the number of HBM modules increases, total memory bandwidth increases but capacity becomes small.

HMA Configuration



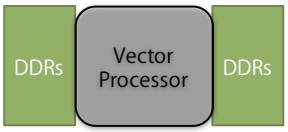
HBM:2048GB/s =14.22GB/s/64GF core
DDR: 153.6GB/s =1.06GB/s/64GF core

HBM Configuration



3072GB/s= 21.33GB/s/64GF core

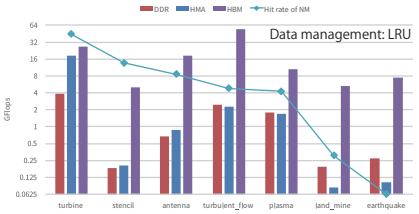
DDR Configuration



512GB/s =3.55GB/s/64GF core

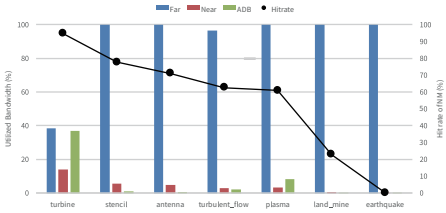
Preliminary Simulations and Future Work

Performance



- As the hit rate to NM increases, the performance of HMA becomes close to the HBM configurations.
- The performance gaps between the HBM and HMA configurations are still large.

Utilized Bandwidth



- Almost all the benchmarks saturate FM bandwidth. Even in this hierarchical configurations, FM becomes bandwidth bottleneck.

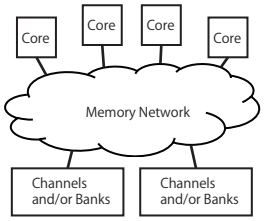
Future Work

Examine more sophisticated data management strategies on HMA for vector supercomputers.

A Simulator of Multi-core Vector Supercomputers

Background and Problem

Current vector processor consists of multiple



Access patterns of applications from multiple cores affects the resources conflicts on a memory network, channels, and banks.

Simulating memory resource conflicts among multiple cores is important !

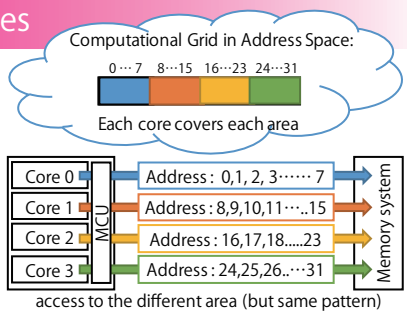
Multi-core Simulation using Pseudo Cores

Pseudo core

A core that issues only memory accesses

→Simple implementation of cores realizes faster simulation speed

→Memory resource conflicts can be simulated by issuing memory accesses from pseudo cores



access to the different area (but same pattern)

(URL) https://www.cal.is.tohoku.ac.jp/
(E-mail) {ryosuke.sato.r4, hikaru.takayashiki.p5}@dc.tohoku.ac.jp