Toward Latency-Aware Data Arrangement on Many-Core Processors

Extended Abstract

Tomoya Yuki
Tokyo Institute of Technology
yuki.t.ab@m.titech.ac.jp

Toshio Endo
Tokyo Institute of Technology
endo@is.titech.ac.jp

KEYWORDS

memory, many-core, latency

1 INTRODUCTION

To achieve higher performance and larger capacity on recent and future architectures, memory devices including 3D stacked memory as HBM and non-volatile memory (NVM) have been integrated. Also structures of memory hierarchy are explored[2].

We are investigating impacts on memory performance by another feature, latency variances caused by placement of processor cores in a chip. In many-core era, dozens or hundreds cores are packed in a 2D chip or chips, and memory controllers are often on the edges of chips. On Intel Xeon Phi processors, cores compose mesh networks, which introduces variances of memory access latency among cores. These variances will be relatively increased in future architecture with reduced latency by stacking processor chips and memory chips.

Our objectives are to understand these impacts on applications performance on upcoming memory hierarchy, towards development of latency-aware data arrangement technologies.

2 PRELIMINARY EXPERIMENTS ON A XEON PHI PROCESSOR

As preliminary experiments towards the above objectives, we have measured memory access latency using a Xeon Phi 7285 processor (72cores, 1.3GHz). The processor is connected with 6 DDR4 DRAM modules and 8 MCDRAM modules. By selecting SNC4 cluster mode in BIOS, 8 MCDRAM modules are separated into 4 numa nodes, each of which has 2 modules. Thus we can fix the target memory modules in measurement easily. Furthermore, we investigated how memory addresses are interleaved between 2 modules.

We have measured memory access latency from every core to memory on a MCDRAM module. In the measurement, effects of prefetching are excluded. As a result, we observe 15.5us (9.6%) difference between the fastest core and the slowest one, which is reproducible.

Although we do not have information on the physical placement of each core, we can presume it with the measured results. Figure1 shows the presumed placement of cores with a heat map colored by the latency. Grey tiles are disabled cores. In the figure, we assume the access target MCDRAM module is put near cores 4 and 5. It shows that we can place cores so that they conform a graduation in the heat map.

We have also measured for other MCDRAM modules and DDR4 modules, and observed a similar tendency. Especially, when we use another MCDRAM module, different cores become fastest according to placement of the memory controller.

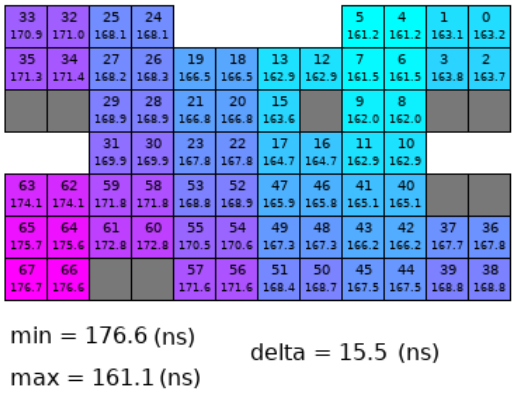

Figure 1: Latency from a MCDRAM module to each core on Xeon Phi Processor

3 FUTURE DIRECTIONS

We demonstrated that the physical location of processor cores affect memory performance. In future, the impacts will be increased with larger number of cores and/or stacking technology of processors and memory chips. Under this assumption, our future directions include the followings:

- Development of emulation technology to assess the impact of core locations on future architectures and devices. The approach of ORNL Siena[2] is highly helpful, but supports of core locations have to be added.
- Development of latency-aware data arrangement technologies like [1] to improve application performance on future architectures.

ACKNOWLEDGMENTS

This research is partly supported by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] Ananya Muddukrishna, Artur Podobas, Mats Brorsson, and Vladimir Vlassov. 2013. Task Scheduling on Manycore Processors with Home Caches. In *Proceedings of Euro-Par'12*. Springer-Verlag, Berlin, Heidelberg, 357–367. https://doi.org/10.1007/978-3-642-36949-0_39

[2] Ivy B. Peng and Jeffrey S. Vetter. 2018. Siena: Exploring the Design Space of Heterogeneous Memory Systems. In *Proceedings of SC '18*. IEEE Press, Piscataway, NJ, USA, Article 33, 14 pages. https://doi.org/10.1109/SC.2018.00036