Toward Latency-Aware Data Arrangement on Many-Core Processors

Tomoya Yuki¹, Toshio Endo¹ (¹Tokyo Institute of Technology)

Background

Recently, many processors which have a complex structure are coming out

- They are composed of unique structure: <u>unique placement</u> of cores and memory controllers
- They have many processor cores. Future architectures will have dozens or handreds cores.
- They and the whole systems using them may have a <u>heterogeneous memory hierarchy</u> composed of Cache, HPM (MCDRAM/HBM), DDR, NVM, SSD, HDD.

Intel Xeon Phi KNM [1]







Latency Checker

We made a latency checker tool for our preliminary experiments. It works on not only standard Xeon processors and also Xeon Phi and AMD EPYC processors.

Latency Checker functions

- ✓ Consider Cacheline size
- ✓ Linked List
- \checkmark Random access mode

memory channel 1

✓ Separate memory interleaving

To separate memory interleaving, the tool use stride access



0x7ffec93e3296

0x7ffec93e3360



Do **Latenty variance** caused by placement of cores and memory modules affect performance?

- Farther cores from a memory controller have high latency, nearer ones have low latency? If so, <u>how much?</u>
- How should we treat memory hierarchies properly? Interleaving many modules gets high bandwidth but on the other hand it causes high latency.
- What should we do to get the best performance on such processors or systems?

[1] <u>https://www.anandtech.com/show/9802/supercomputing-15-intels-knights-landing-xeon-phi-silicon-on-display</u> [2] https://www.techspot.com/news/81350-amd-unveils-2nd-gen-epyc-cpus-world-fastest.html [3] https://pc.watch.impress.co.jp/docs/column/kaigai/1202957.html [4] https://www.nextplatform.com/2019/11/13/a64fx-arm-chip-gets-a-big-push-from-cray/

memory channel 2 Can **randomize** memory address for preventing memory memory channel 1 from prefetch 0x7ffec93e3296 memory channel 2

Preliminary Experiments

Intel Xeon Phi KNM

Toward our above objectives, we have measured memory access latency from every core to memory. We presumed physical placement of cores with measured result.

All2Al	Il mode	All2All and Quadrant m lge and border cores have	odes, high latency							
MCDRAM	DDR4 an	d cores around center ha	ve low one.	Expected location						
53 52 23 22 24 25 21 20 1 0 170.2 170.2 168.4 168.4 168.4 168.5 168.4 170.1 170.3 61 60 31 30 3 2 51 50 29 28 7 6 169.8 169.8 168.0 167.6 167.6 167.6 168.1 168.0 169.7 169.8	53 52 23 22 147.5 147.5 143.8 143.8 143.8 143.8 144.0 144.0 147.5 147.5 147.5 147.5 147.5 147.5 147.5 147.5 147.7 143.7 143.7 143.8 12 51 50 29 28 7 6 143.7 143.7 143.7 143.6 143.7 143.7 143.4 143.5 143.7			of a MCDRAM module						
39 38 9 8 59 37 36 37 36 167.6 167.6 167.2 167.2 167.2 167.2 167.2 167.7 167.8 <td>39 38 9 8 59 37 36 143.8 143.9 143.6 143.6 143.8 143.6 143.7 143.6 143.7 143.7 143</td> <td>33 170.9 1</td> <td>SNC4 mod 32 25 24 71.0 168.1 168.1</td> <td>le MCDRAM 5 4 1 0 161.2 161.2 163.1 163.2</td>	39 38 9 8 59 37 36 143.8 143.9 143.6 143.6 143.8 143.6 143.7 143.6 143.7 143.7 143	33 170.9 1	SNC4 mod 32 25 24 71.0 168.1 168.1	le MCDRAM 5 4 1 0 161.2 161.2 163.1 163.2						
min = 170.317558 max = 167.061677 delta = 3.255881 (ns)	min = 147.669091 max = 143.260333 delta = 4.408758 (ns)	35 171.3 1	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	3 13 12 7 6 3 2 .5 162.9 162.9 161.5 161.5 163.8 163.7 0 15 9 8 5 163.8 163.7						
Ouadra	nt mode		168.9 168.9 166.8 166 31 30 23 22	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$						
MCDRAM DDR4			169.9 169.9 167.8 167	.8 164.7 164.7 162.9 162.9						
53 52 23 22 166.3 21 20 1 0 0 53 52 23 22 22 20 1 0 166.3 167.7 167.6 165.3 165.5 165.5 165.5 165.5 165.5 165.5 165.5 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4 165.4		63 174.1 1 65 175.7 1 67 176.7 1	62 59 58 53 52 74.1 171.8 171.8 168.8 168 64 61 60 55 54 75.6 172.8 172.8 170.5 170 66 57 56 171.6 171	2 47 46 41 40						
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$		min = 176.6 (ns) max = 161.1 (ns)	delta = 15.5 (ns)						
SNC4 mod	e DDR4	In SNC4 mode,								
33 32 25 24 137.4 137.5 137.4 137.4		we can see a beauti	ful gradation from	om the nearest core to farthest one						
35 34 27 26 19 18 13 137.4 142 142 142 142 143 143 143 144 144 144 144 144 144 144 144 144 144 144 144 144 144 144 144 144 144	3 12 7 6 3 2 .7 142.6 148.3 148.3 154.4 154.4	on not only MCDR	AM but also DDR	R4.						
29 28 21 20 15 137.4 137.4 137.5 137.4 140	9 8 145.5 145.6	Environment								
31302322137137.4137.4137.4137.4137.4137	16 11 10 .3 137.3 144.4 144.4	Processor	Intel [®] Xeon Phi™ Proc	essor 7285 (Kinghts Mill)						
63625958535247137.4137.4137.4137.4137.4137.4137.4	46 41 40 137.4 142.6 142.7	Num of cores	68							
65 64 61 60 55 54 49 137.4 1	48 43 42 37 36 .4 137.4 145.4 145.4 148.3 148.3	Memory	MCDRAM*8, DDR4*6							
67 66 57 56 51 137.4 137.4 137.4 137.4 141	L 50 45 44 39 38 .8 142.0 145.5 145.5 151.0 151.3	Numa mode	All2All, Quadrant, SNC	24						
		Conditions	Disabled h/w prefetch	, Separated memory interleaving(SNC4)						
min = 154.6 (ns) max = 137.2 (ns)	elta = 17.4 (ns)	Buffer Size	100 MiB							

AMD EPYC Rome

0x7ffec93e3232

AMD EPYC Rome can be set NPS4 mode which is like SNC4 mode on Xeon Phi. We measured latencies on a machine which has 2 sockets for the processors. So it appeared 8 numa nodes. Each of them has 1/8 of total cores and memory





8	9	10	11		24	25	26	27	72	73	74	75		88	89	90	91	
141.5	140.7	145.9	141.5		150.3	150.0	151.2	151.3	270.2	269.8	270.6	270.3		274.6	273.7	274.8	274.8	
12	13	14	15		28	29	30	31	76	77	78	79		92	93	94	95	
140.9	144.2	142.0	141.5		150.3	150.2	151.2	151.3	270.0	270.2	270.2	270.3		273.7	273.9	274.5	274.6	
				-														
				_														
32	33	34	35		48	49	50	51	96	97	98	99		112	113	114	115	
163.9	163.2	164.5	164.6		170.6	170.2	170.8	170.8	250.5	250.4	251.0	250.9		275.7	276.3	276.7	276.5	
36	37	38	39		52	53	54	55	100	101	102	103		116	117	118	119	
163.5	163.4	164.8	164.6		170.6	170.4	171.0	170.8	250.3	250.5	251.6	250.7		276.0	275.8	276.7	276.5	
40	41	42	43		56	57	58	59	104	105	106	107		120	121	122	123	
163.6	163.3	164.7	164.6		170.4	170.4	170.8	170.9	250.2	250.6	250.8	250.7		276.5	276.2	276.4	276.7	
44	45	46	47		60	61	62	63	108	109	110	111		124	125	126	127	
163.7	163.3	164.6	164.6		170.5	170.2	170.9	170.8	250.3	250.6	251.5	250.7		275.7	275.8	276.7	276.3	
											*(color	inσ a	re do	ne ii	n eac	h soc	ke
min	min = 170.9 (ns)						min = 286.1 (ns)											
	delta = 30.2 (ns)						delta = 35.8 (ns)											
max	= 14	06	(ns)	C			\	- /	max	= 25	02	nc)	C			(-,	
ΠUΛ	17		(13)						mux	— ZJ	0.2	(15)						

Access across sockets costs a lot (110ns).

But the varience of latencies is observed even in the same socket.

AMD EPYC [™] 7702 Processor (EPYC Rome)
64 cores * 2 socket = 128
DDR4
NPS4 (4 numa nodes)
Ramdom access
100 MiB

Future Directions

Development of **latency-aware data arrangement technologies**

Prior research: Task Scheduling on Manycore Processors with Home Caches[5]





We predict these latencies will increase when the number of cores increase. Cache coherency does not scale. We have to consider how to treat dozens or handreds cores.

[5] 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

[6] 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

This work is partly supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO)

