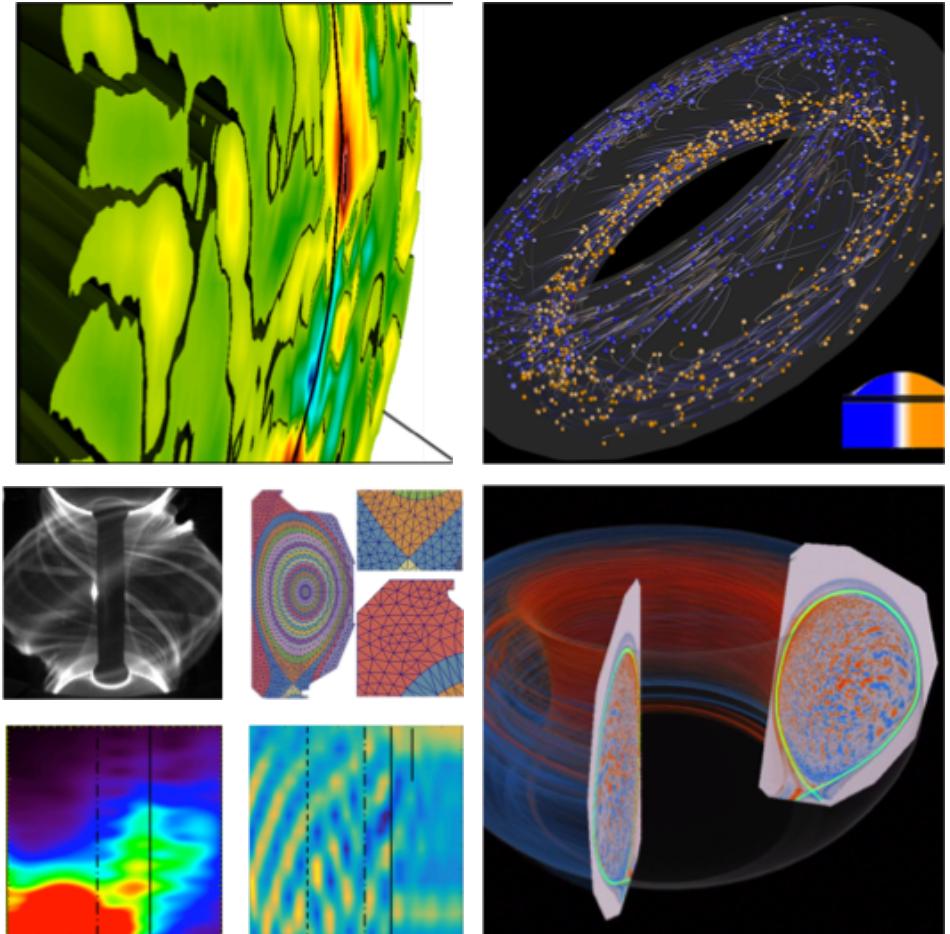


LIKWID at NERSC



Charlene Yang

Application Performance Group
cjyang@lbl.gov

What is LIKWID



- ‘Like I Knew What I’m Doing’ -- Erlangen Regional Computing Center
- A toolset:

likwid-topology	node topology
likwid-pin	process/thread affinity
likwid-memsweeper	cleanup memory & LLC
likwid-powermeter	power measurements
likwid-setFrequencies	CPU/uncore frequency manipulation
likwid-perfctr	hardware counter measurements
likwid-mpirun	hardware counter + MPI
likwid-bench	micro-benchmarking
likwid-agent	system monitoring
likwid-genTopoCfg	generate and store topology file

- Marker API with C/C++, Fortran90

likwid-topology



```
-----  
CPU name:      Intel(R) Xeon Phi(TM) CPU 7250 @ 1.40GHz  
CPU type:      Intel Xeon Phi (Knights Landing) (Co)Processor  
CPU stepping:  1  
*****  
Hardware Thread Topology  
*****  
Sockets:        1  
Cores per socket: 68  
Threads per core: 4  
-----  


| HWThread | Thread | Core | Socket | Available |
|----------|--------|------|--------|-----------|
| 0        | 0      | 0    | 0      | *         |
| 1        | 0      | 1    | 0      | *         |
| 2        | 0      | 2    | 0      | *         |
| 3        | 0      | 3    | 0      | *         |
| 4        | 0      | 4    | 0      | *         |
| 5        | 0      | 5    | 0      | *         |
| 6        | 0      | 6    | 0      | *         |
| 7        | 0      | 7    | 0      | *         |
| 8        | 0      | 8    | 0      | *         |
| 9        | 0      | 9    | 0      | *         |
| 10       | 0      | 10   | 0      | *         |
| 11       | 0      | 11   | 0      | *         |
| 12       | 0      | 12   | 0      | *         |
| 13       | 0      | 13   | 0      | *         |
| 14       | 0      | 14   | 0      | *         |


```

likwid-topology



```
*****
Cache Topology
*****  
Level: 1  
Size: 32 kB  
Cache groups: ( 0 68 136 284 ) ( 1 69 137 285 ) ( 2 70 138 286 ) ( 3 71 139 287 ) ( 4 72 140 288 ) ( 5 73 141 289 ) ( 6 74 142 210 ) ( 7 75 143 211 ) ( 8 76 144 212 ) ( 9 77 145 213 ) ( 10 78 146 214 ) ( 11 79 147 215 ) ( 12 80 148 216 ) ( 13 81 149 217 ) ( 14 82 150 218 ) ( 15 83 151 219 ) ( 16 84 152 220 ) ( 17 85 153 221 ) ( 18 86 154 222 ) ( 19 87 155 223 ) ( 20 88 156 224 ) ( 21 89 157 225 ) ( 22 90 158 226 ) ( 23 91 159 227 ) ( 24 92 160 228 ) ( 25 93 161 229 ) ( 26 94 162 230 ) ( 27 95 163 231 ) ( 28 96 164 232 ) ( 29 97 165 233 ) ( 30 98 166 234 ) ( 31 99 167 235 ) ( 32 100 168 236 ) ( 33 101 169 237 ) ( 34 102 170 238 ) ( 35 103 171 239 ) ( 36 104 172 240 ) ( 37 105 173 241 ) ( 38 106 174 242 ) ( 39 107 175 243 ) ( 40 108 176 244 ) ( 41 109 177 245 ) ( 42 110 178 246 ) ( 43 111 179 247 ) ( 44 112 180 248 ) ( 45 113 181 249 ) ( 46 114 182 250 ) ( 47 115 183 251 ) ( 48 116 184 252 ) ( 49 117 185 253 ) ( 50 118 186 254 ) ( 51 119 187 255 ) ( 52 120 188 256 ) ( 53 121 189 257 ) ( 54 122 190 258 ) ( 55 123 191 259 ) ( 56 124 192 260 ) ( 57 125 193 261 ) ( 58 126 194 262 ) ( 59 127 195 263 ) ( 60 128 196 264 ) ( 61 129 197 265 ) ( 62 130 198 266 ) ( 63 131 199 267 ) ( 64 132 200 268 ) ( 65 133 201 269 ) ( 66 134 202 270 ) ( 67 135 203 271 )  
-----  
Level: 2  
Size: 1 MB  
Cache groups: ( 0 68 136 284 1 69 137 285 ) ( 2 70 138 286 3 71 139 287 ) ( 4 72 140 288 5 73 141 289 ) ( 6 74 142 210 7 75 143 211 ) ( 8 76 144 212 9 77 145 213 ) ( 10 78 146 214 11 79 147 215 ) ( 12 80 148 216 13 81 149 217 ) ( 14 82 150 218 15 83 151 219 ) ( 16 84 152 220 17 85 153 221 ) ( 18 86 154 222 19 87 155 223 ) ( 20 88 156 224 21 89 157 225 ) ( 22 90 158 226 23 91 159 227 ) ( 24 92 160 228 25 93 161 229 ) ( 26 94 162 230 27 95 163 231 ) ( 28 96 164 232 29 97 165 233 ) ( 30 98 166 234 31 99 167 235 ) ( 32 100 168 236 33 101 169 237 ) ( 34 102 170 238 35 103 171 239 ) ( 36 104 172 240 37 105 173 241 ) ( 38 106 174 242 39 107 175 243 ) ( 40 108 176 244 41 109 177 245 ) ( 42 110 178 246 43 111 179 247 ) ( 44 112 180 248 45 113 181 249 ) ( 46 114 182 250 47 115 183 251 ) ( 48 116 184 252 49 117 185 253 ) ( 50 118 186 254 51 119 187 255 ) ( 52 120 188 256 53 121 189 257 ) ( 54 122 190 258 55 123 191 259 ) ( 56 124 192 260 57 125 193 261 ) ( 58 126 194 262 59 127 195 263 ) ( 60 128 196 264 61 129 197 265 ) ( 62 130 198 266 63 131 199 267 ) ( 64 132 200 268 65 133 201 269 ) ( 66 134 202 270 67 135 271 )  
-----  
*****  
NUMA Topology  
*****  
NUMA domains: 1  
-----  
Domain: 0  
Processors: ( 0 68 136 284 1 69 137 285 2 70 138 286 3 71 139 287 4 72 140 288 5 73 141 289 6 74 142 210 7 75 143 211 8 76 144 212 9 77 145 213 10 78 146 214 11 79 147 215 12 80 148 216 13 81 149 217 14 82 150 218 15 83 151 219 16 84 152 220 17 85 153 221 18 86 154 222 19 87 155 223 20 88 156 224 21 89 157 225 22 90 158 226 23 91 159 227 24 92 160 228 25 93 161 229 26 94 162 230 27 95 163 231 28 96 164 232 29 97 165 233 30 98 166 234 31 99 167 235 32 100 168 236 33 101 169 237 34 102 170 238 35 103 171 239 36 104 172 240 37 105 173 241 38 106 174 242 39 107 175 243 40 108 176 244 41 109 177 245 42 110 178 246 43 111 179 247 44 112 180 248 45 113 181 249 46 114 182 250 47 115 183 251 48 116 184 252 49 117 185 253 50 118 186 254 51 119 187 255 52 120 188 256 53 121 189 257 54 122 190 258 55 123 191 259 56 124 192 260 57 125 193 261 58 126 194 262 59 127 195 263 60 128 196 264 61 129 197 265 62 130 198 266 63 131 199 267 64 132 200 268 65 133 201 269 66 134 202 270 67 135 203 271 )  
Distances: 10  
Free memory: 93294.1 MB  
Total memory: 96563.2 MB  
-----
```

likwid-pin



- **likwid-pin -c N:0,8,16,24 ./xthi.x**
- **likwid-pin -c S0:0,8@S1:0,8 ./xthi.x**

HSW

```
Hello from rank 0, thread 0, on nid00028. (core affinity = 0)
Hello from rank 0, thread 1, on nid00028. (core affinity = 8)
Hello from rank 0, thread 2, on nid00028. (core affinity = 16)
Hello from rank 0, thread 3, on nid00028. (core affinity = 24)
```

- **likwid-pin -c E:N:128:2:4 ./xthi.x**

KNL

```
Hello from rank 0, thread 0, on nid02308. (core affinity = 0)
Hello from rank 0, thread 1, on nid02308. (core affinity = 68)
Hello from rank 0, thread 2, on nid02308. (core affinity = 1)
Hello from rank 0, thread 3, on nid02308. (core affinity = 69)
* snip *
Hello from rank 0, thread 126, on nid02308. (core affinity = 63)
Hello from rank 0, thread 127, on nid02308. (core affinity = 131)
```

- **likwid-perfctr takes the same specification as its processor list**

LIKWID for profiling



- **likwid-perfctr (threaded) + likwid-mpirun (MPI/hybrid)**
 - no GUI
 - low overhead
 - no code instrumentation required
 - no root access required
 - no extra modules required to be installed
 - > SDE, VTune, etc
 - > CrayPat-tracing
 - > VTune
 - > VTune
- use Linux '**msr**' module to access MSR (Model Specific Register) files
- Cori:
 - module load vtune
 - sbatch/salloc **--perf=vtune**
 - module load likwid

**May change in the future
e.g. --perf=likwid**

likwid-perfctr -a



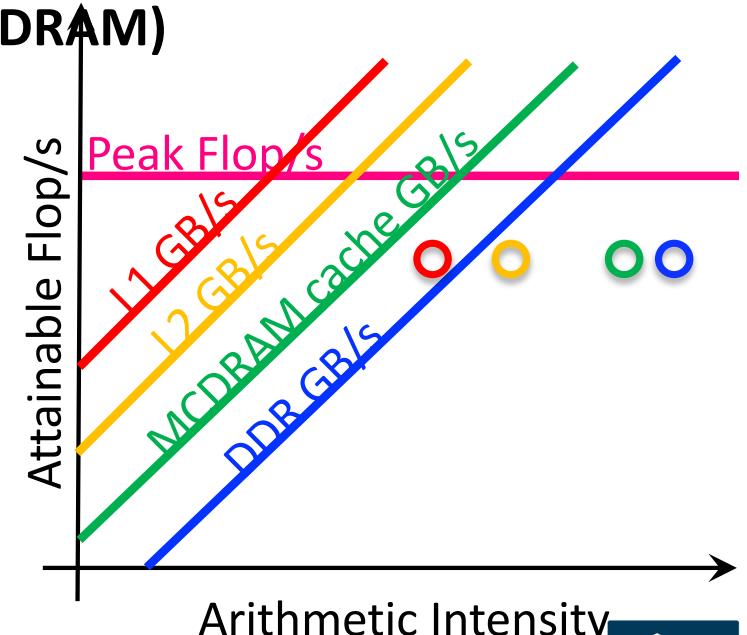
- performance groups on KNL

Group name	Description
<hr/>	
HBM_OFFCORE	Memory bandwidth in MBytes/s for High Bandwidth Memory (HBM)
TLB_INSTR	L1 Instruction TLB miss rate/ratio
FLOPS_SP	Single Precision MFLOP/s
BRANCH	Branch prediction miss rate/ratio
L2CACHE	L2 cache miss rate/ratio
ENERGY	Power and Energy consumption
FRONTEND_STALLS	Frontend stalls
ICACHE	Instruction cache miss rate/ratio
TLB_DATA	L2 data TLB miss rate/ratio
MEM	Memory bandwidth in MBytes/s
DATA	Load to store ratio
L2	L2 cache bandwidth in MBytes/s
FLOPS_DP	Double Precision MFLOP/s
CLOCK	Power and Energy consumption
HBM_CACHE	Memory bandwidth in MBytes/s for High Bandwidth Memory (HBM)
HBM	Memory bandwidth in MBytes/s for High Bandwidth Memory (HBM)
UOPS_STALLS	UOP retirement stalls

Data collection



- GPP kernel from BerkeleyGW
- Arithmetic Intensity = $\text{FLOPS} / \text{Bytes}$ (= SDE / VTune)
- $= \text{FLOPS/sec} / \text{Bytes/sec}$
- $= \text{FLOPS_DP} / \text{Bandwidth}$
- AI (DRAM) = $\text{FLOPS_DP} / \text{Bandwidth (DRAM)}$
- AI (MCDRAM)= $\text{FLOPS_DP} / \text{Bandwidth (MCDRAM)}$
- AI (L2) = $\text{FLOPS_DP} / \text{Bandwidth (L2)}$
- AI (L1) = $\text{FLOPS_DP} / \text{Bandwidth (L1)}$
- Performance = FLOPS_DP



FLOPS/sec



- GPP kernel on KNL: **171.960 GFLOPS/sec**
 - UOPS_RETIRED_PACKED SIMD
 - UOPS_RETIRED_SCALAR SIMD
- likwid-perfctr -C 0-63 -g **FLOPS_DP** ./gpp.knl.ex 512 2 32768 20
 - 8*UOPS_RETIRED_PACKED SIMD+UOPS_RETIRED_SCALAR SIMD

Metric	Sum	Min	Max	Avg
Runtime (RDTSC) [s] STAT	940.8064	14.7001	14.7001	14.7001
Runtime unhalted [s] STAT	402.9130	6.2371	9.8444	6.2955
Clock [MHz] STAT	96000.0155	1499.9955	1500.0007	1500.0002
CPI STAT	86.0772	1.3396	1.5850	1.3450
DP MFLOP/s (SSE assumed) STAT	44456.2105	688.9334	729.9324	694.6283
DP MFLOP/s (AVX assumed) STAT	86957.6422	1347.4354	1429.2337	1358.7132
DP MFLOP/s (AVX512 assumed) STAT	171960.5065	2664.4393	2827.8362	2686.8829
Packed MUOPS/s STAT	21250.7102	329.2510	349.6506	332.0424
Scalar MUOPS/s STAT	1954.7786	30.4313	30.6312	30.5434

DRAM/MCDRAM bandwidth



- GPP kernel on KNL: **DDR 2.59GB/s + MCDRAM 63.71GB/s**
 - MC_CAS_READS/ MC_CAS_WRITES
 - EDC_RPQ_INSERTS/ EDC_WPQ_INSERTS
 - EDC_MISS_CLEAN/ EDC_MISS_DIRTY
- likwid-perfctr -C 0-63 -g **HBM_CACHE** ./gpp.knl.ex 512 2 32768 20

Metric	Sum	Min	Max	Avg
Runtime (RDTSC) [s] STAT	896.4352	14.0068	14.0068	14.0068
Runtime unhalted [s] STAT	390.2173	6.0393	9.6183	6.0971
Clock [MHz] STAT	95979.5220	1499.6763	1499.6807	1499.6800
CPI STAT	83.4239	1.2985	1.5496	1.3035
MCDRAM Memory read bandwidth [MBytes/s] STAT	63246.3054	0	63246.3054	988.2235
MCDRAM Memory read data volume [GBytes] STAT	885.8769	0	885.8769	13.8418
MCDRAM Memory writeback bandwidth [MBytes/s] STAT	468.4857	0	468.4857	7.3201
MCDRAM Memory writeback data volume [GBytes] STAT	6.5620	0	6.5620	0.1025
MCDRAM Memory bandwidth [MBytes/s] STAT	63714.7910	0	63714.7910	995.5436
MCDRAM Memory data volume [GBytes] STAT	892.4389	0	892.4389	13.9444
DDR Memory read bandwidth [MBytes/s] STAT	2569.3065	0	2569.3065	40.1454
DDR Memory read data volume [GBytes] STAT	35.9877	0	35.9877	0.5623
DDR Memory writeback bandwidth [MBytes/s] STAT	21.1772	0	21.1772	0.3309
DDR Memory writeback data volume [GBytes] STAT	0.2966	0	0.2966	0.0046
DDR Memory bandwidth [MBytes/s] STAT	2590.4837	0	2590.4837	40.4763
DDR Memory data volume [GBytes] STAT	36.2843	0	36.2843	0.5669

L2 bandwidth



- GPP kernel on KNL: **L2 96.80GB/s**
 - L2_REQUESTS_REFERENCE
 - OFFCORE_RESPONSE_0_OPTIONS
- likwid-perfctr -C 0-63 -g **L2 ./gpp.knl.ex 512 2 32768 20**

Metric	Sum	Min	Max	Avg
Runtime (RDTSC) [s] STAT	895.5200	13.9925	13.9925	13.9925
Runtime unhalted [s] STAT	392.3078	6.0719	9.6599	6.1298
Clock [MHz] STAT	95999.4279	1499.9861	1499.9914	1499.9911
CPI STAT	83.8844	1.3055	1.5567	1.3107
L2 non-RFO bandwidth [MBytes/s] STAT	96803.9243	1498.7686	1904.3169	1512.5613
L2 non-RFO data volume [GByte] STAT	1354.5272	20.9715	26.6461	21.1645
L2 RFO bandwidth [MBytes/s] STAT	0	0	0	0
L2 RFO data volume [GByte] STAT	0	0	0	0
L2 bandwidth [MBytes/s] STAT	96803.9243	1498.7686	1904.3169	1512.5613
L2 data volume [GByte] STAT	1.354528e+06	20971.5004	26646.1299	21164.4950

L1 bandwidth



- GPP kernel on KNL: **L1 170.77GB/s**
 - MEM_UOPS_RETIRED_ALL_LOADS
 - MEM_UOPS_RETIRED_ALL_STORES
- likwid-perfctr -C 0-63 -g **DATA** ./gpp.knl.ex 512 2 32768 20
 - $(\text{MEM_UOPS_RETIRED_ALL_LOADS} + \text{MEM_UOPS_RETIRED_ALL_STORES}) * 64 / \text{runtime}$
 - -g DATA is for load-to-store ratio, but can be used to estimate L1 bandwidth

Compare with SDE/VTune



- **SDE FLOPS:**
 - sde64 -knl -d -iform 1 -omix my_mix.out -global_region -- ./gpp.knl.ex 512 2 32768 20
 - ./parse-sde.sh my_mix.out
 - --->Total FLOPs = 2775769815463
- **VTune Bytes:**
 - amplxe-cl -collect memory-access -finalization-mode=deferred -r my_vtune/ -- ./gpp.knl.ex 512 2 32768 20
 - amplxe-cl -report summary -r my_vtune/ > my_vtune.summary
 - ./parse-vtune.sh my_vtune.summary
 - DDR --->Total Bytes = 35983553088
 - HBM --->Total Bytes = 963486016448
- <http://www.nersc.gov/users/application-performance/measuring-arithmetic-intensity/>

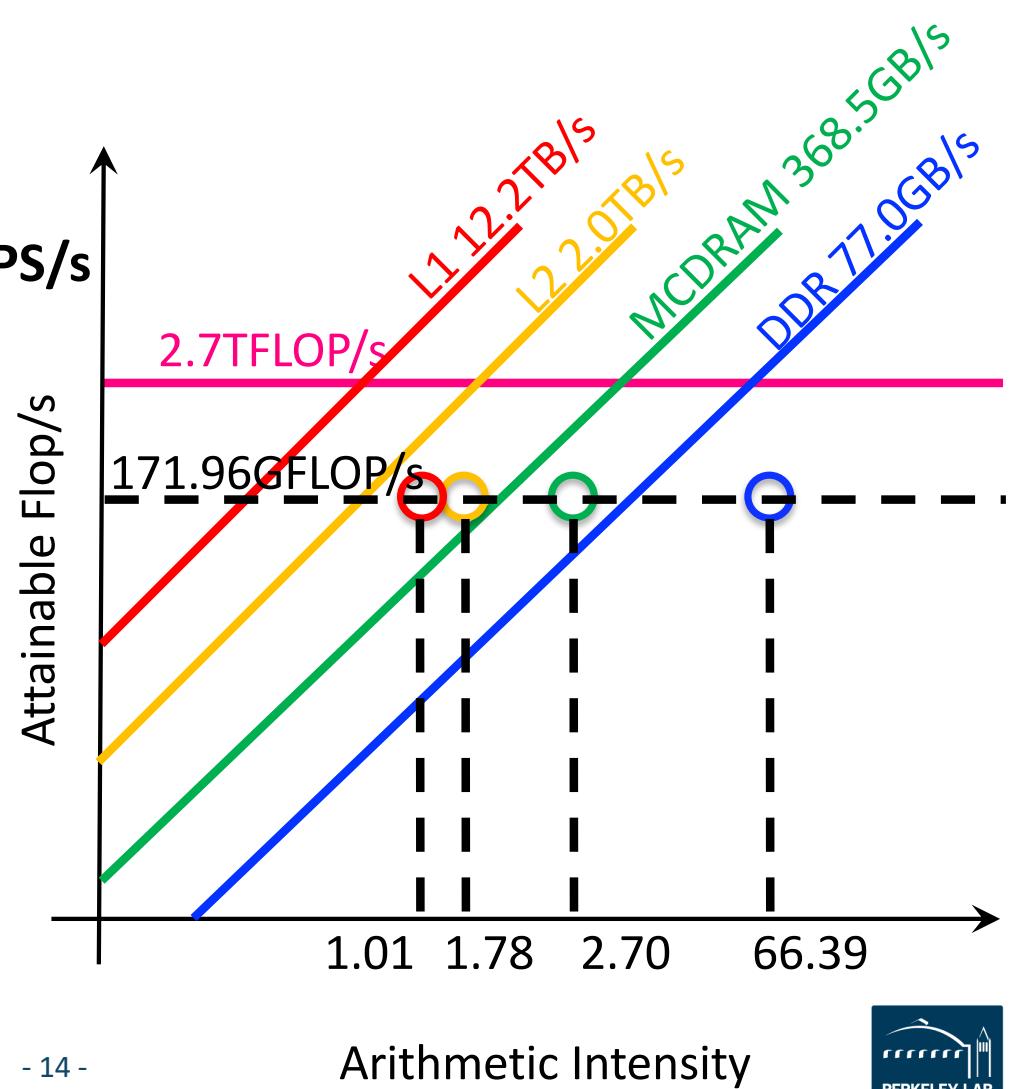
LIKWID:
2527.81 GFLOPS **~8.9%**

LIKWID:
DDR: 36.28 GB **~0.8%**
HBM: 892.44 GB **~7.4%**

Roofline model



- AI (DRAM): 66.39
- AI (MCDRAM): 2.70
- AI (L2): 1.78
- AI (L1): 1.01
- Performance: **171.960 GFLOPS/s**



likwid-mpirun



- **srun -n 2 -c 32 --cpu-bind=cores likwid-perfctr -C 0,8 -g MEM -o test_%h_%p_%r.txt ./xthi.x**
- %h -- hostname, %p -- process ID, %r -- MPI rank
- **likwid-mpirun -pin S0:0,8_S1:0,8 -g MEM ./xthi.x**

HSW

```
Hello from rank 0, thread 0, on nid00191. (core affinity = 0)
Hello from rank 0, thread 1, on nid00191. (core affinity = 8)
Hello from rank 1, thread 0, on nid00191. (core affinity = 16)
Hello from rank 1, thread 1, on nid00191. (core affinity = 24)
```

- **Uncore counters are measured on a per-socket basis**

likwid-perfctr -m



- **cc -fopenmp -DLIKWID_PERFMON -I\$LIKWID_INCLUDE -L\$LIKWID_LIB -llikwid -dynamic test.c -o test.x**
- **likwid-perfctr -C 0-3 -g MEM -m ./test.x**

```
#include <likwid.h>

.....
LIKWID_MARKER_INIT;
#pragma omp parallel {
    LIKWID_MARKER_THREADINIT;
}
#pragma omp parallel {
    LIKWID_MARKER_START("foo");
    #pragma omp for
    for(i = 0; i < N; i++) {
        data[i] = omp_get_thread_num();
    }
    LIKWID_MARKER_STOP("foo");
}
LIKWID_MARKER_CLOSE;
```



focus on specific code regions

Summary



LIKWID

- **node topology, process/thread affinity, micro-benchmarking**
- **performance counters -> roofline model**
 - FLOPS/s and Bytes/s for different levels of cache
 - low overhead, high accuracy

