COMP 322: Fundamentals of Parallel Programming

Lecture 33: Task Affinity with Places

Vivek Sarkar
Department of Computer Science, Rice University
vsarkar@rice.edu

https://wiki.rice.edu/confluence/display/PARPROG/COMP322

Worksheet #32: MPI Gather

```
1.   MPI.Init(args) ;
2.   int myrank = MPI.COMM_WORLD.Rank() ;
3.   int numProcs = MPI.COMM_WORLD.Size() ;
4.   int size = ...;
5.   int[] sendbuf = new int[size];
6.   int[] recvbuf = new int[???];
7.   . . . // Each process initializes sendbuf
8.   MPI.COMM_WORLD.Gather(sendbuf, 0, size, MPI.INT,
9.                         recvbuf, 0, size, MPI.INT,
10.                        0/*root*/);
11.  . . .
12.  MPI.Finalize();
```

Question: In the space below, indicate what values should be provided instead of ??? in line 6, and why.

Answer:

recvbuf should be allocated with numProcs*size elements for Gather. Since recvbuf also needs to be allocated in the root, line 6 can be replaced by:

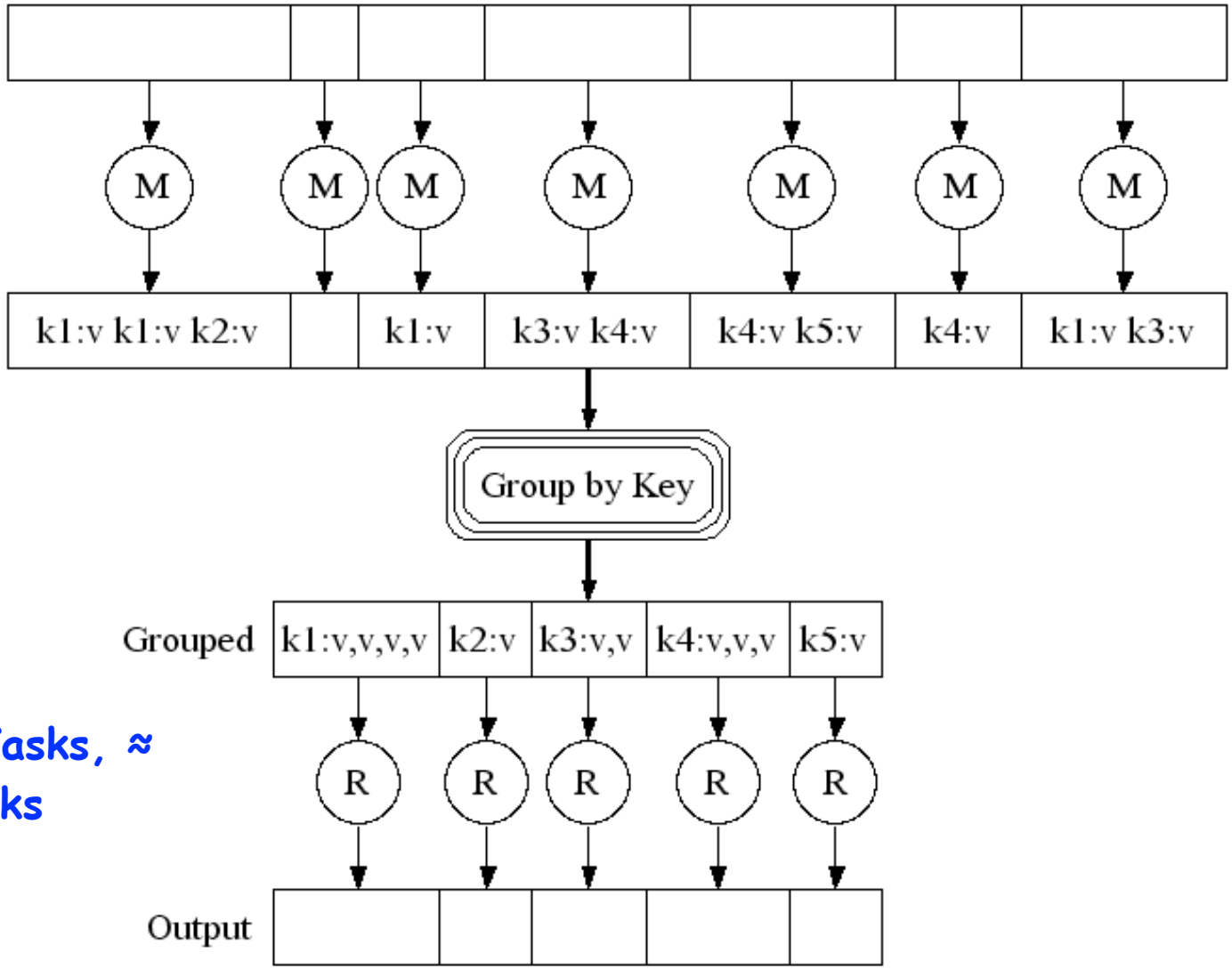
```
6.int[] recvbuf = (myrank==0) ? new int[numProcs*size] : null;
```


MapReduce Execution (Recap from Lecture 8)

Fine granularity tasks: many more map tasks than machines

Bucket sort to get same keys together

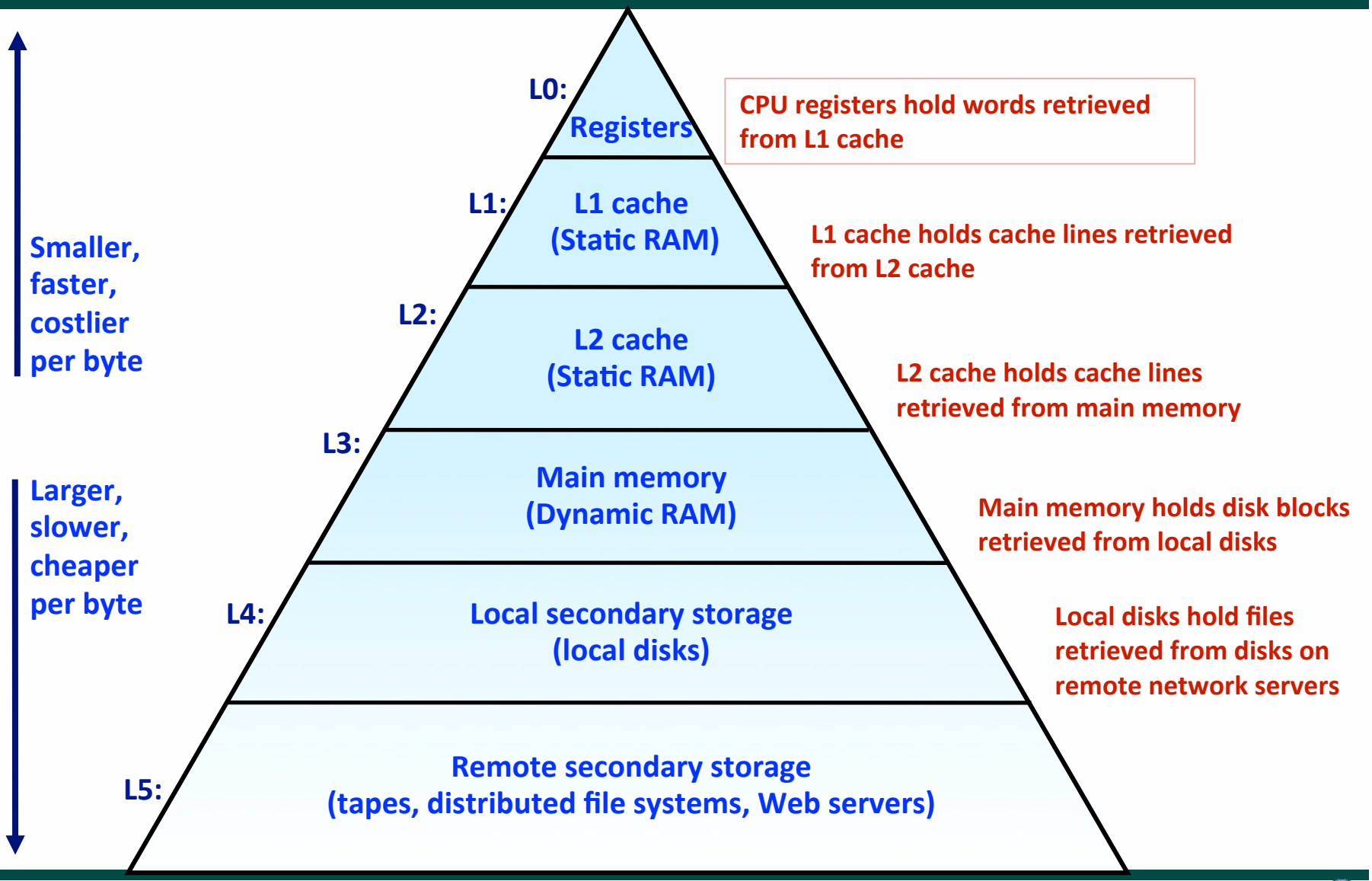
2000 servers => ≈ 200,000 Map Tasks, ≈ 5,000 Reduce tasks



PseudoCode for WordCount
(Recap from Lecture 8)

```
1.  map(String input_key, String input_value):
2.    // input_key: document name
3.    // input_value: document contents
4.    for each word w in input_value:
5.      EmitIntermediate(w, "1"); // Produce count of words
6.
7.  reduce(String output_key, Iterator intermediate_values):
8.    // output_key: a word
9.    // intermediate_values: a list of counts
10.    int result = 0;
11.    for each v in intermediate_values:
12.      result += ParseInt(v); // get integer from key-value
13.    Emit(AsString(result));
```


An example Memory Hierarchy --- what is the cost of a Memory Access?



Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

Storage Trends

SRAM

Metric	1980	1985	1990	1995	2000	2005	2010	2010:1980
$/MB	19,200	2,900	320	256	100	75	60	320
access (ns)	300	150	35	15	3	2	1.5	200

DRAM

Metric	1980	1985	1990	1995	2000	2005	2010	2010:1980
$/MB	8,000	880	100	30	1	0.1	0.06	130,000
access (ns)	375	200	100	70	60	50	40	9
typical size (MB)	0.064	0.256	4	16	64	2,000	8,000	125,000

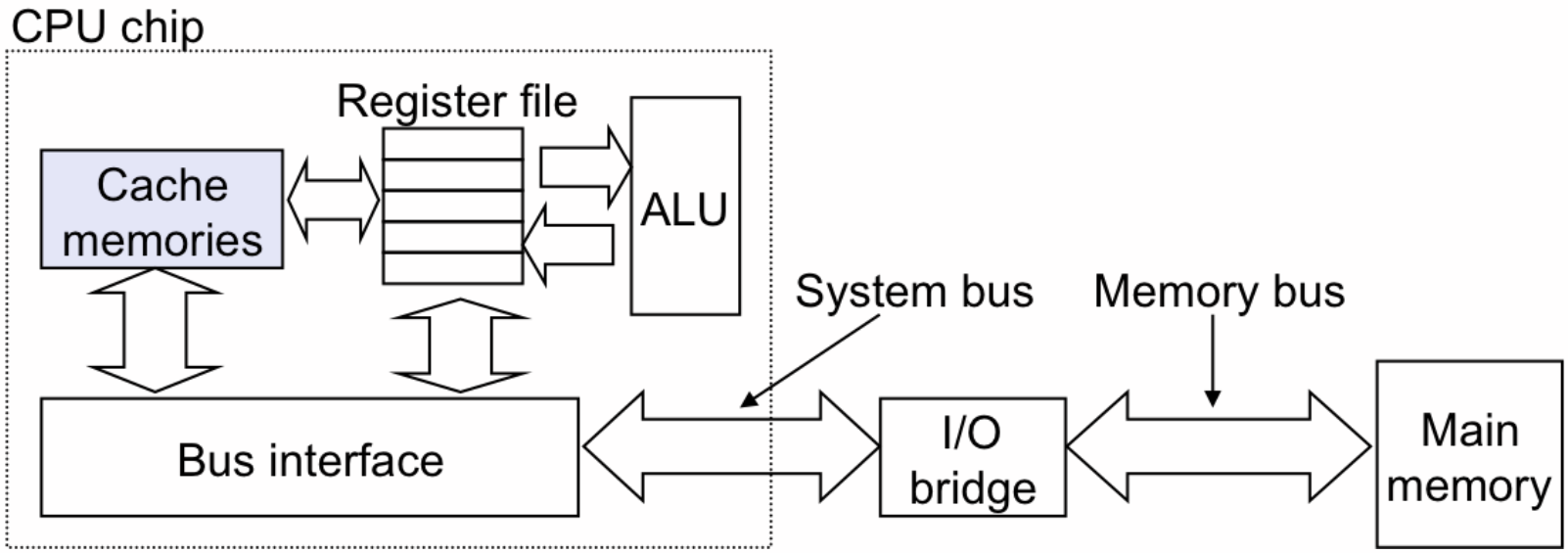
Disk

Metric	1980	1985	1990	1995	2000	2005	2010	2010:1980
$/MB	500	100	8	0.30	0.01	0.005	0.0003	1,600,000
access (ms)	87	75	28	10	8	4	3	29
typical size (MB)	1	10	160	1,000	20,000	160,000	1,500,000	1,500,000

Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

Cache Memories

- **Cache memories are small, fast SRAM-based memories managed automatically in hardware.**
  - **—Hold frequently accessed blocks of main memory**
- **CPU looks first for data in caches (e.g., L1, L2, and L3), then in main memory.**
- **Typical system structure:**



Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

Examples of Caching in the Hierarchy

Hierarchy Level	d by
Registers	
TLB	
L1 cach	
L2 cac	
Virtua	
Buffer	
Disk ca	
Network	
Browser cac	
Web cache	y server

Ideally one would desire an indefinitely large memory capacity such that any particular ... word would be immediately available. ... We are ... forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.

A. W. Burks, H. H. Goldstine, and J. von Neumann
Preliminary Discussion of the Logical Design of an Electronic Computing Instrument (1946)

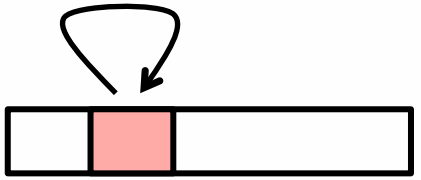
Ultimate goal: create a large pool of storage with average cost per byte that approaches that of the cheap storage near the bottom of the hierarchy, and average latency that approaches that of fast storage near the top of the hierarchy.

Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

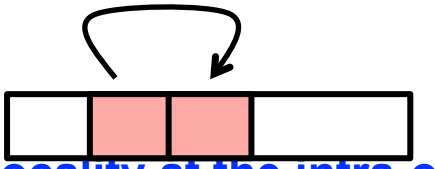
Locality

- **Principle of Locality:**
  - —Empirical observation: Programs tend to use data and instructions with addresses near or equal to those they have used recently

- **Temporal locality:**
  - —Recently referenced items are likely to be referenced again in the near future



- **Spatial locality:**
  - —Items with nearby addresses tend to be referenced close together in time
  - —A Java programmer can only influence spatial locality at the intra-object level
    - The garbage collector and memory management system determines inter-object placement



Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

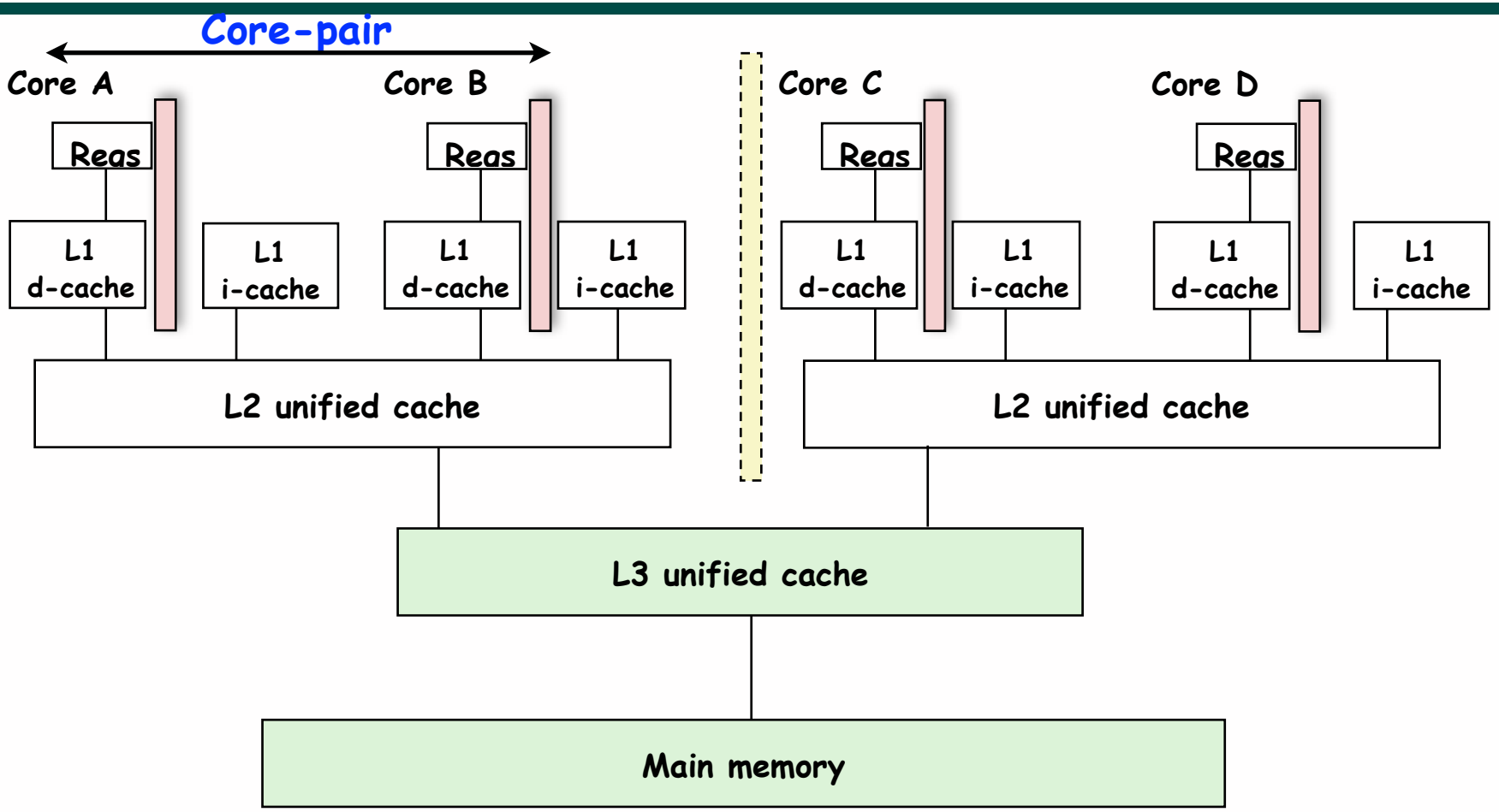
Locality Example

```
sum = 0;
for (i = 0; i < n; i++)
    sum += a[i];
return sum;
```

- **Data references**
  - —Reference array elements in succession (stride-1 reference pattern). — Spatial locality
  - —Reference variable `sum` each iteration. — Temporal locality
- **Instruction references**
  - —Reference instructions in sequence. — Spatial locality
  - —Cycle through loop repeatedly. — Temporal locality

Source: http://www.cs.cmu.edu/afs/cs/academic/class/15213-f10/www/lectures/09-memory-hierarchy.pptx

Memory Hierarchy in a Multicore Processor



- Memory hierarchy for a single Intel Xeon Quad-core E5440 HarperTown processor chip
  - —A SUG@R node contains TWO such chips, for a total of 8 cores

Programmer Control of Task Assignment to Processors

- The parallel programming constructs that we've studied thus far result in tasks that are assigned to processors *dynamically* by the HJ runtime system
  - —Programmer does not worry about task assignment details
- Sometimes, programmer control of task assignment can lead to significant performance advantages due to improved locality
- Motivation for HJ "places"
  - —Provide the programmer a mechanism to restrict task execution to a subset of processors for improved locality
  - —Current HJlib implementation supports one level of locality via places, but future HJlib versions will support hierarchical places

Places in HJ

HJ programmer defines mapping from HJ tasks to set of places

HJ runtime defines mapping from places to one or more worker Java threads per place

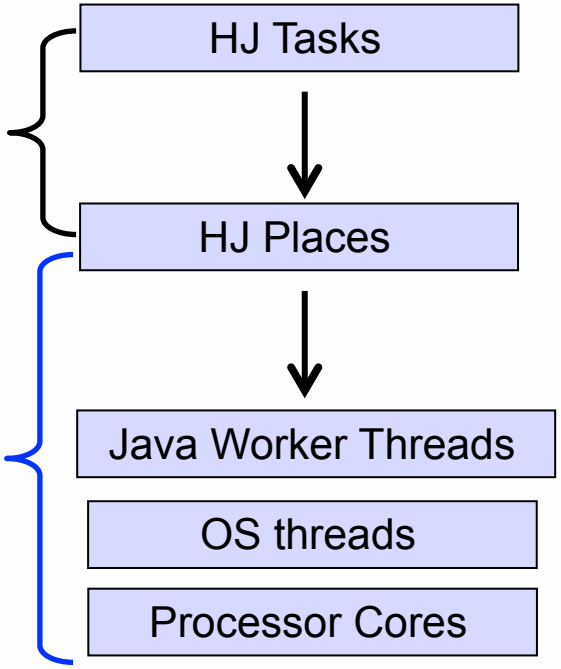
The option

```
HjSystemProperty.numPlaces.setProperty(p);
HjSystemProperty.numWorkers.setProperty(w);
```

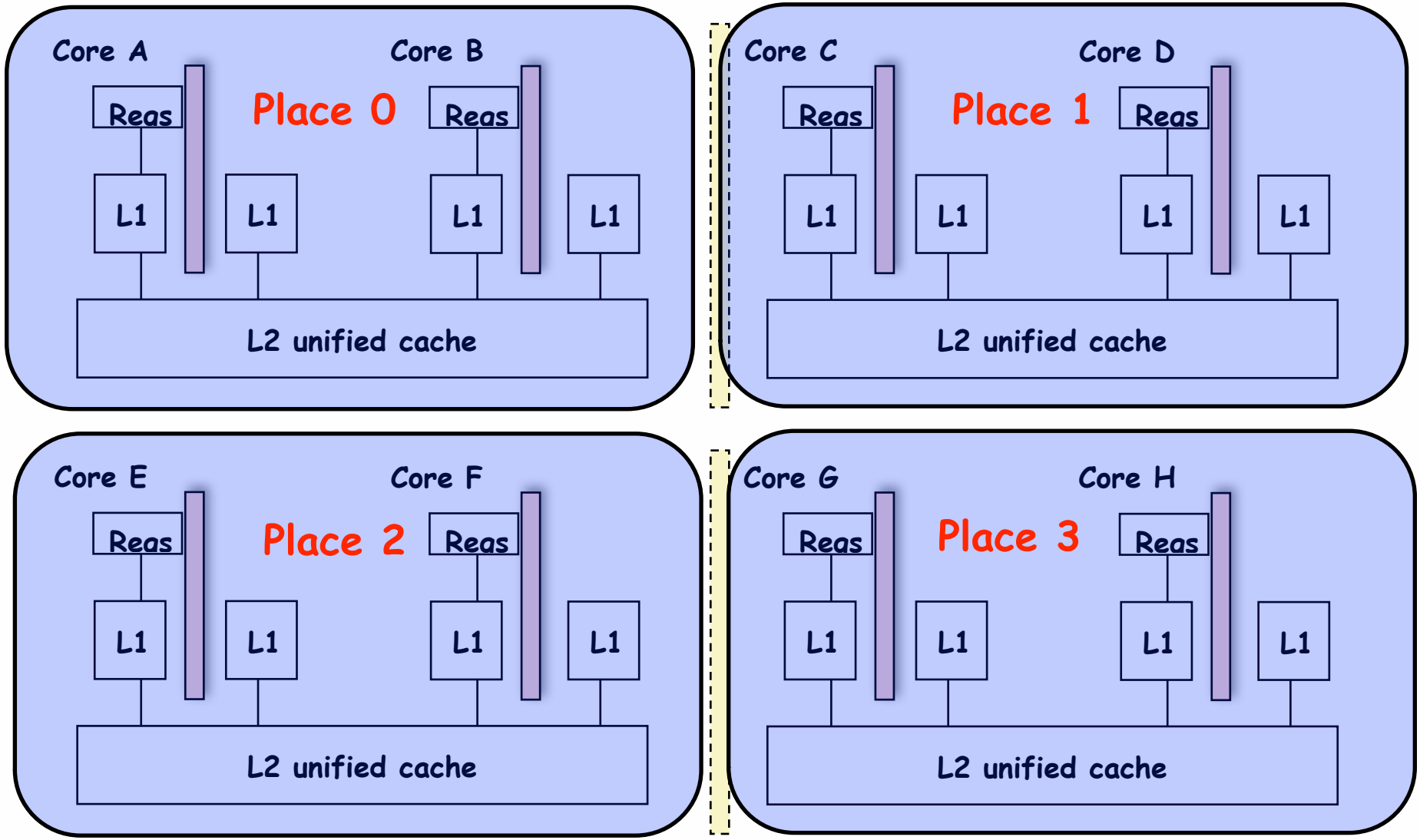
when executing an HJ program can be used to specify

- p, the number of places
- w, the number of worker threads per place

we will abbreviate this as p:w



Example of 4:2 option on an 8-core node (4 places w/ 2 workers per place)



Places in HJlib

here() = place at which current task is executing

numPlaces() = total number of places (runtime constant)

Specified by value of **p** in runtime option:

```
HjSystemProperty.numPlaces.setProperty(p);
```

place(i) = place corresponding to index i

<place-expr>.toString() returns a string of the form "place(id=0)"

<place-expr>.id() returns the id of the place as an int

asyncAt(P, () -> S)

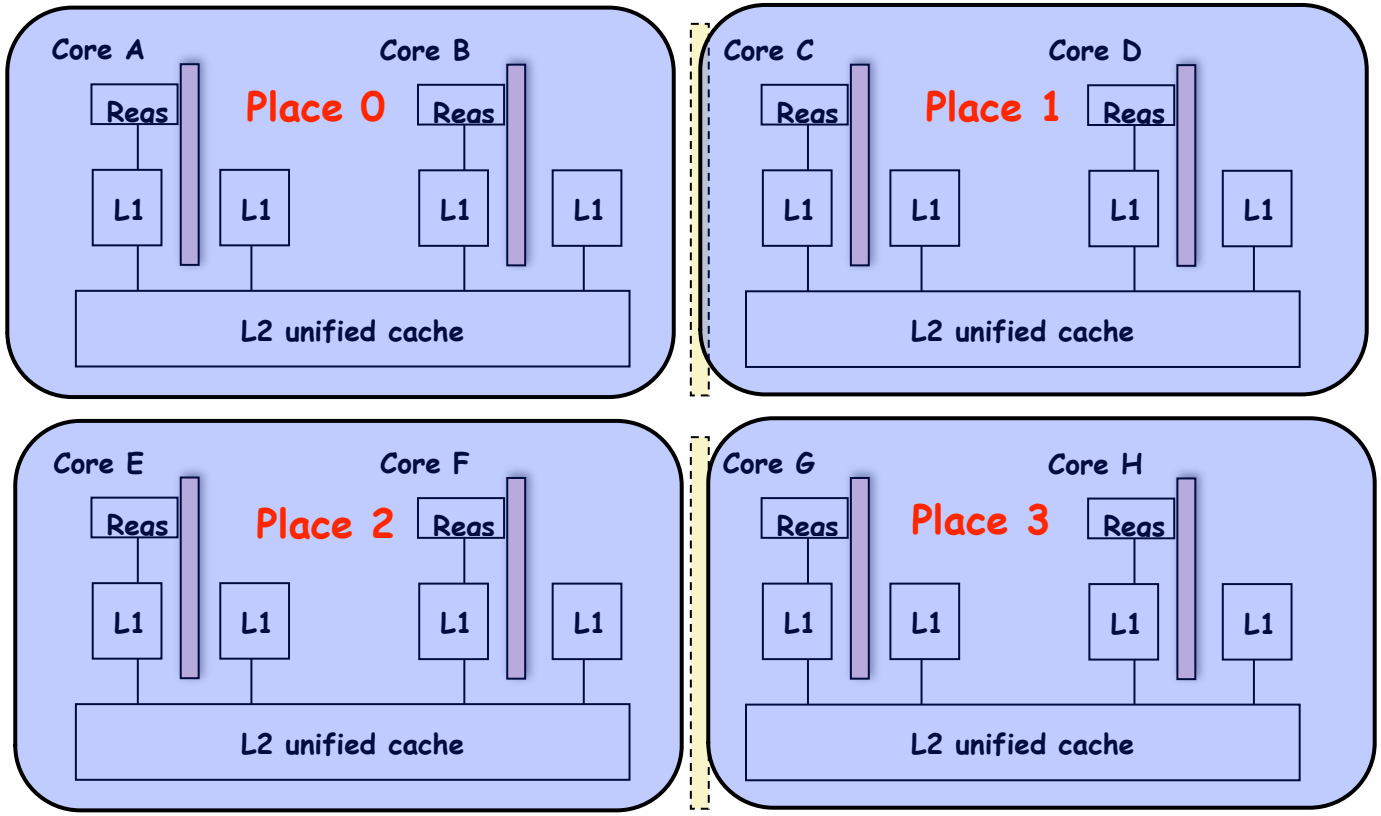
- Creates new task to execute statement S at place P
- **async(() -> S)** is equivalent to **asyncAt(here(), () -> S)**
- Main program task starts at **place(0)**

Note that **here()** in a child task refers to the place P at which the child task is executing, not the place where the parent task is executing

Example of 4:2 option on an 8-core node (4 places w/ 2 workers per place)

```
// Main program starts at place 0
asyncAt(place(0), () -> S1);
asyncAt(place(0), () -> S2);
```

```
asyncAt(place(1), () -> S3);
asyncAt(place(1), () -> S4);
asyncAt(place(1), () -> S5);
```

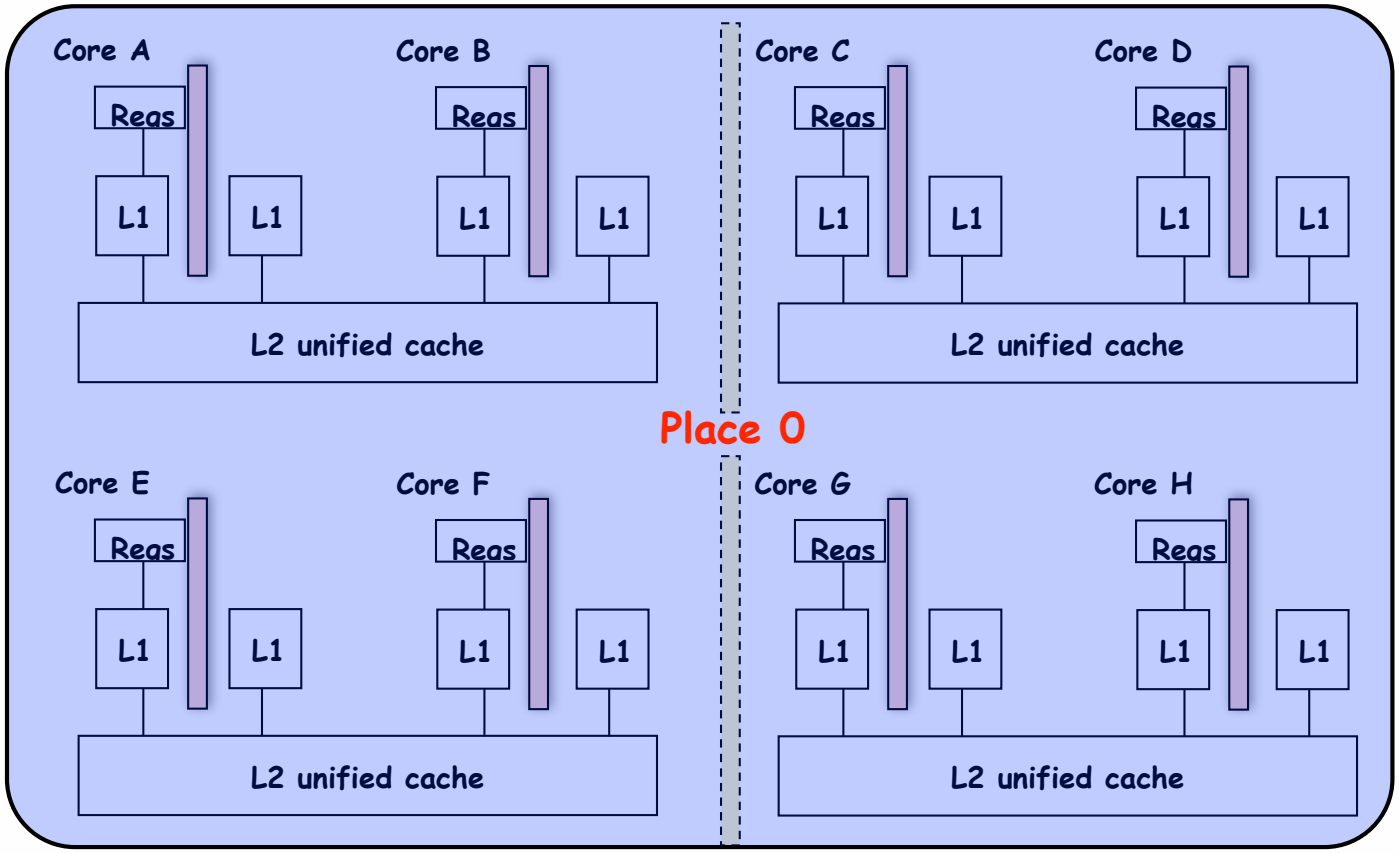



```
asyncAt(place(2), () -> S6);
asyncAt(place(2), () -> S7);
asyncAt(place(2), () -> S8);
```

```
asyncAt(place(3), () -> S9);
asyncAt(place(3), () -> S10);
```


Example of 1:8 option (1 place w/ 8 workers per place)

All async's run at place 0 when there's only one place!



HJ program with places (pseudocode)

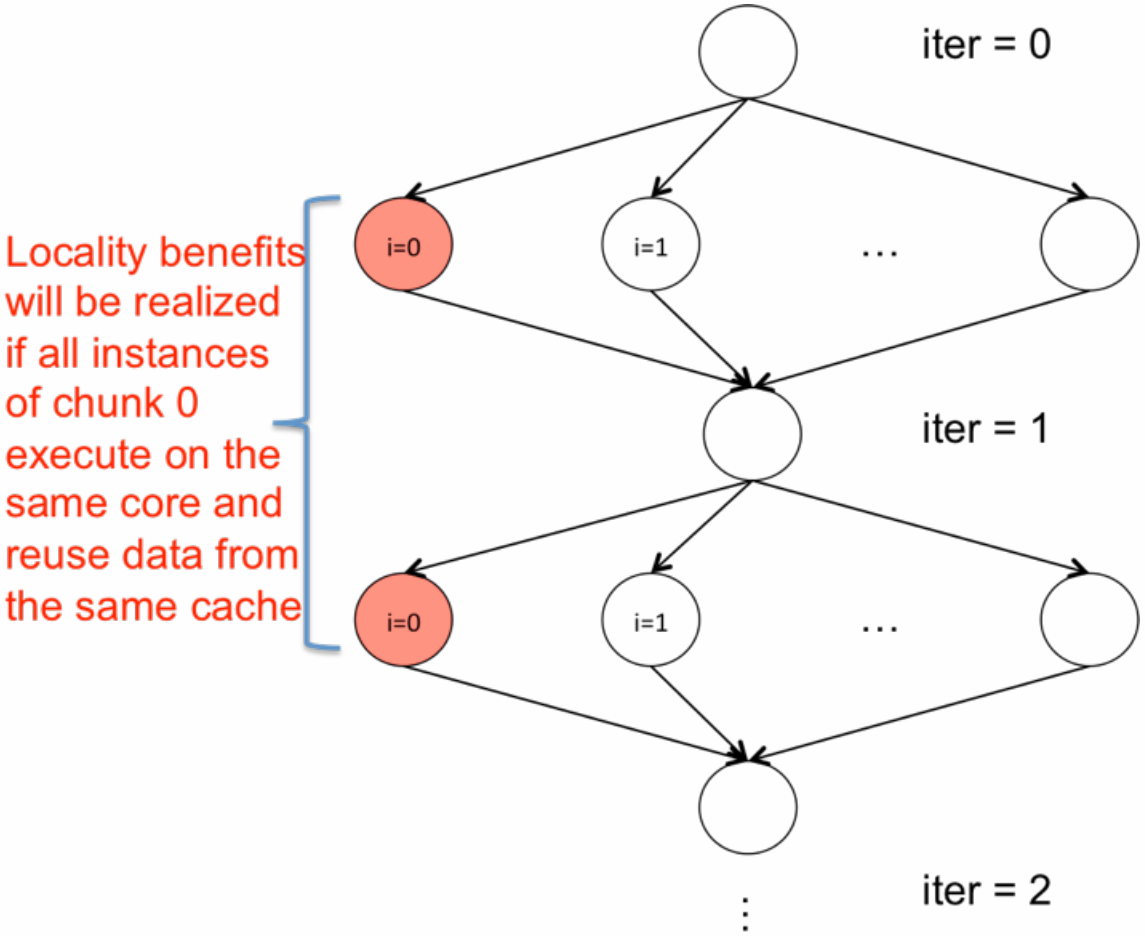
```
1  class T1 {
2    final place affinity;
3    . . .
4    // T1's constructor sets affinity to place where instance was created
5    T1() { affinity = here; ... }
6    . . .
7  }
8  . . .
9  finish { // Inter-place parallelism
10   System.out.println("Parent place =", here); // Parent task s place
11   for (T1 a = . . .) {
12     async at (a.affinity) { // Execute async at place with affinity to a
13       a.foo();
14       System.out.println("Child place =", here); // Child task's place
15     } // async
16   } // for
17 } // finish
18 . . .
```


Chunked Fork-Join Iterative Averaging Example with Places

```
1.  public void runDistChunkedForkJoin(
2.      int iterations, int numChunks, Dist dist) {
3.    // dist is a user-defined map from int to HjPlace
4.    for (int iter = 0; iter < iterations; iter++) {
5.      finish(() -> {
6.        forseq (0, numChunks - 1, (jj) -> {
7.          asyncAt(dist.get(jj), () -> {
8.            forseq (getChunk(1, n, numChunks, jj), (j) -> {
9.              myNew[j] = (myVal[j-1] + myVal[j+1]) / 2.0;
10.            }
11.         });
12.       });
13.     });
14.      double[] temp = myNew; myNew = myVal; myVal = temp;
15.   } // for iter
16. }
```

- Chunk jj is always executed in the same place for each iter
- Method runDistChunkedForkJoin can be called with different values of distribution parameter d

Analyzing Locality of Fork-Join Iterative Averaging Example with Places



Worksheet #33: impact of distribution on parallel completion time (instead of locality)

Name: _____ Netid: _____

```
1.  public void sampleKernel(
2.      int iterations, int numChunks, Dist d) {
3.    for (int iter = 0; iter < iterations; iter++) {
4.      finish(() -> {
5.        forseq (0, numChunks - 1, (jj) -> {
6.          asyncAt(dist.get(jj), () -> {
7.            perf.doWork(jj);
8.            // Assume that time to process chunk jj = jj units
9.          });
10.       });
11.     });
12.      double[] temp = myNew; myNew = myVal; myVal = temp;
13.   } // for iter
14. } // sample kernel
```

- Assume an execution with n places, each place with one worker thread
- Will a block or cyclic distribution for d have a smaller abstract completion time, assuming that all tasks on the same place are serialized?