

# Compiler Support for Software Cache Coherence

Sanket Tavarageri  
The Ohio State University  
Email: tavarageri.1@osu.edu

Wooil Kim      Josep Torrellas  
University of Illinois at Urbana-Champaign  
Email: {kim844,torrella}@illinois.edu

P Sadayappan  
The Ohio State University  
Email: sadayappan.1@osu.edu

**Abstract**—The advent of multi-core processors with a large number of cores and heterogeneous architecture poses challenges for achieving scalable cache coherence. Several recent research efforts have focused on simplifying or abandoning hardware cache coherence protocols. However, this adds a significant burden on the programmer, unless automated compiler support is developed.

In this paper, we develop compiler support for parallel systems that delegate the task of maintaining cache coherence to software. Algorithms to automatically insert software cache coherence instructions into parallel applications are presented. This frees the programmer from having to manually insert coherence annotations, which can be tedious and error-prone. Experimental evaluation over a number of benchmarks demonstrates that effective compiler techniques can make software cache coherence competitive with hardware coherence schemes both in terms of energy and performance.

## I. INTRODUCTION

As the number of transistors on a chip continues to scale, multi-core and many-core processors are expected to have hundreds of cores in the near future. The use of a hierarchy of caches and/or scratchpad memory is universal, to enable fast access to spatially and temporally reused data. But this also introduces the cache coherence problem — keeping shared variables coherent across multiple processors.

Commodity multicore processors currently enforce cache coherence through snooping-based or directory-based protocols. However, snoopy protocols [2] rely on the existence of a shared bus to enforce cache coherence, and therefore are not appropriate when a network-on-chip is used to connect different processing elements. Network-on-chip is crucial to achieve scalable performance for architectures that have a large number of cores. Hence, directory-based hardware coherence mechanisms [19] have been proposed for large-scale parallel computers. Directory schemes however impose additional storage requirements and increase memory access latency, potentially causing significant network traffic due to directory maintenance operations. Despite progress made towards addressing the aforementioned problems [14], [13], significant challenges remain for hardware cache coherence schemes [28]. An additional challenge with hardware-based coherence schemes is their attendant complexity: hardware coherence protocols are known to be notoriously complex to design and verify [1].

Consequently, there is a drive to simplify hardware cache coherence protocols or to do away with them. For example, Runnemed from the DARPA UHPC program [3] provides

scratchpads and software-managed incoherent caches, shifting the responsibility of coherence to the software. The Cohesion system [16] proposes a hybrid memory model that combines software and hardware coherence schemes — for regular applications, coherence is software-driven, while hardware coherence is invoked for irregular code. The 48-core Intel SCC research many-core processor [21] drops hardware cache coherence altogether, and instead defines a new memory type to facilitate communication between cores. A few other research projects [6], [15] have proposed simplifying the implementation of hardware cache coherence protocols by restricting the types of sharing patterns that the application can exhibit. In some cases, the programmer/compiler can help identify potentially stale data for self-invalidation, further increasing performance.

Therefore, it is of considerable interest to develop complementary compiler algorithms to automatically insert software coherence instructions into application programs, so that valid execution can be ensured on systems that do not provide hardware coherence support, but expect the compiler (or programmer) to ensure coherence.

The topic of compiler-directed cache coherence was previously addressed by several efforts (e.g., [7]). Even though that had shown promise, software cache coherence was discontinued mainly for a couple of reasons: 1) the kind of programs for which the compiler support was available was limited and therefore, it placed a lot of responsibility on the shoulders of the programmer 2) snoopy-based hardware coherence protocols performed well on the parallel computers which had a small number of cores. But the very high degree of parallelism in upcoming multiprocessor systems, such as those proposed in the DARPA UHPC and DOE XStack programs, is prompting a renewed interest in alternatives to the traditional hardware-based coherence schemes. In this paper, we develop compiler analyses for efficient software managed cache coherence. We take a multi-pronged approach using several strategies:

- 1) For memory accesses with affine index expressions, we develop compile time analyses to precisely mark only those variables for invalidation that could become stale in the private cache due to other writes from other cores. Similarly, we also develop analyses to precisely identify the data to be written back from the private caches to shared caches because other cores might need to access the data in future epochs.
- 2) For repetitive irregular (non-affine) accesses, we present *inspector*-based schemes that exactly demarcate data for

coherence management.

- 3) Other irregular parallel applications are handled via methods that write back and invalidate data across synchronization points and preserve data locality of read-only data.

The paper makes the following contributions:

- Compiler algorithms to automatically instrument parallel applications with cache management instructions that write back and invalidate cached data,
- Efficient compiler analyses using the Polyhedral-model [12] to precisely identify data for software-based cache coherence for affine computations,
- Compiler analyses using the inspector-executor paradigm to maintain cache coherence for iterative irregular computations,
- Experimental demonstration using a range of programs demonstrating that the developed compiler-based techniques are competitive with hardware coherence schemes in terms of performance and energy consumption, at a lower hardware cost.

## II. OVERVIEW & BACKGROUND

The software orchestrates cache coherence using the following coherence primitives.

- **Writeback:** The address of a variable is specified in the instruction and if the addressed location exists in the private cache and has been modified, then it is written to a shared cache or main memory.
- **Invalidate:** The instruction causes any cached copy of the variable in the private cache to be discarded (*self-invalidation*) so that the next read to the variable fetches data from the shared cache.

if processor A has to send an updated value of a shared variable X to processor B, then processor A issues a *writeback* instruction on X, and processor B later invalidates X so that a subsequent read to X fetches the latest value from the shared cache.

Fig. 1 shows the API for the *invalidate* and *writeback* instructions.

### A. Execution Model

**Release Consistency:** The execution of parallel programs consists of *epochs* (intervals between global synchronization points). Examples of epochs include, code executed between successive *barriers*, the code region between acquiring and releasing of a *lock*. In an epoch, data which was written

```

invalidate_word(void *addr);
invalidate_dword(void *addr);
invalidate_qword(void *addr);
invalidate_range(void *addr, int num_bytes);

writeback_word(void *addr);
writeback_dword(void *addr);
writeback_qword(void *addr);
writeback_range(void *addr, int num_bytes);

```

Fig. 1: Coherence API list

potentially by other cores in previous epochs and that a core may need to read in the current epoch are *invalidated*. Before the end of the epoch, all the data that a core has written in the current epoch and that may be needed by other processors in future epochs are *written-back* to the shared level cache.

Before an epoch completes, all prior memory operations, including ordinary load/store instructions and coherence instructions, are completed. Then the next epoch can start, and the subsequent memory operations can be initiated. Further, ordering constraints between memory instructions are respected: The order of a store to address *i* and the subsequent writeback for address *i* should be preserved in the instruction pipeline of the processor and caches. Similarly, the order of invalidation to address *j* and a subsequent load from address *j* should be preserved in the pipeline and caches to guarantee fetching of the value from the shared cache.

**Coherence Operations at Cache line granularity:** Coherence operations are performed at the granularity of cache lines — all the lines that overlap with specified addresses are invalidated or written-back. If the specified data are not present in cache, then coherence instructions have no effect. In addition, *writeback instructions write back only modified words of the line*. In doing so, writeback instructions avoid the incorrectness issue that may arise from false sharing: if two processors are writing to variables that get mapped to the same cache line, and whole cache lines (and not just the *dirty* words) are written-back, then one processor’s modified words may be overwritten with another processor’s clean words. Therefore, per-word *dirty* bits are used to keep track of words of a cache line that are modified.

### B. Notation

The code shown in Fig. 2 is used as a working example to illustrate the notation and the compiler algorithm in the next section.

**Sets:** A set *s* is defined as:

$$s = \{[x_1, \dots, x_m] : c_1 \wedge \dots \wedge c_n\}$$

where each  $x_i$  is a tuple variable and each  $c_j$  is a constraint. The iteration spaces of statements can be represented as sets. For example, the iteration space of statement S1 in the code shown in Fig. 2 can be specified as the set  $I^{S_1}$ :  $I^{S_1} = \{S1[t_1, t_2, t_3] : (0 \leq t_1 \leq tsteps - 1) \wedge (0 \leq t_2 \leq n - 1) \wedge (1 \leq t_3 \leq n - 1)\}$

**Relations:** A relation *r* is defined as:

$$r = \{[x_1, \dots, x_m] \mapsto [y_1, \dots, y_n] : c_1 \wedge \dots \wedge c_p\}$$

where each  $x_i$  is an input tuple variable, each  $y_j$  is an output tuple variable and each  $c_k$  is a constraint. Array accesses

```

for (t1=0; t1 <= tsteps - 1; t1++) {
  #pragma omp parallel for private(t3)
  for (t2=0; t2 <= n - 1; t2++) {
    for (t3=1; t3 <= n - 1; t3++) {
      S1: B[t2][t3] = B[t2][t3+1] + 1;
    }
  }
}

```

Fig. 2: A loop nest

appearing in the code may be modeled as relations from iteration spaces to access functions of the array references. The two accesses to array ‘B’ in Fig. 2,  $B[t2][t3]$  and  $B[t2][t3+1]$ , are represented as the following relations:

$$r_{write}^{S_1} = \{S1[t_1, t_2, t_3] \mapsto B[t'_2, t'_3] : (t'_2 = t_2) \wedge (t'_3 = t_3)\}$$

$$r_{read}^{S_1} = \{S1[t_1, t_2, t_3] \mapsto B[t'_2, t'_3] : (t'_2 = t_2) \wedge (t'_3 = t_3 + 1)\}$$

**The Apply Operation:** The apply operation on a relation  $r$  and a set  $s$  produces a set  $s'$  denoted by,  $s' = r(s)$  and is mathematically defined as:

$$(\vec{x} \in s') \iff (\exists \vec{y} \text{ s.t. } \vec{y} \in s \wedge (\vec{y} \mapsto \vec{x}) \in r)$$

The set of array elements accessed by an array reference in a loop (data-footprint) may be derived by *applying* access function *relations* on the iteration space *sets*. For the array accesses in the example code shown in Fig. 2, data-footprints of the two accesses are:  $r_{write}^{S_1}(I^{S_1}), r_{read}^{S_1}(I^{S_1})$ .

**The Inverse Operation:** The inverse operation  $r = r_k^{-1}$  operates on a relation  $r_k$  to produce a new relation  $r$  such that  $r$  has the same constraints as  $r_k$  but with the input and output tuple variables swapped.  $(\vec{x} \mapsto \vec{y} \in r) \iff (\vec{y} \mapsto \vec{x} \in r_k)$ .

### C. Polyhedral Dependences

In the Polyhedral model, for affine computations, dependence analysis [11] can precisely compute flow (Read After Write - RAW) and output (Write After Write - WAW) dependences between dynamic instances of statements. The dependences are expressed as maps from source iterations to target iterations involved in the dependence.

The flow dependence determined by polyhedral dependence analysis (for example, using ISL [26]) for the code in Fig. 2 is:

$$\mathcal{D}_{flow} = \{S1[t_1, t_2, t_3] \mapsto S1[t_1 + 1, t_2, t_3 - 1] : (0 \leq t_1 \leq tsteps - 2) \wedge (0 \leq t_2 \leq n - 1) \wedge (2 \leq t_3 \leq n - 1)\}$$

The relation characterizes the flow dependence that exists between the write reference  $B[t2][t3]$  and the read reference  $B[t2][t3+1]$ . An analysis tool like ISL can also be used to emit information regarding *live-in* data: data that are read in the loop but are not produced by any statement instances in the scope of analysis. A list containing maps from an iteration point that reads live-in data to the live-in array elements that it reads is computed. For the running example, the live-in maps are:

$$\mathcal{D}_{live-in} = \left. \begin{aligned} &\{S1[0, t_2, t_3] \mapsto B[t_2, t_3 + 1] : 0 \leq t_2 \leq n - 1 \wedge 1 \leq t_3 \leq n - 2; \\ &S1[t_1, t_2, n - 1] \mapsto B[t_2, n] : 0 \leq t_1 \leq tsteps - 1 \wedge 0 \leq t_2 \leq n - 1 \} \end{aligned} \right\}$$

The two maps capture live-in data read for read reference  $B[t2][t3+1]$ .

## III. COMPILER OPTIMIZATION FOR REGULAR CODE

The iteration space of an epoch in a parallel loop is modeled by considering iterator values of the parallel loop and its surrounding loops as parameters. In the parallel loop in Fig. 2, the  $t2$  loop is parallel and an iteration of  $t2$  constitutes a *parallel task* executed in an epoch. Its iteration space is modeled by considering values of iterators  $t1$  and  $t2$  as parameters -  $t_p$  and  $t_q$  respectively:  
 $I_{current}^{S_1} = \{S1[t_1, t_2, t_3] : (t_1 = t_p) \wedge (t_2 = t_q) \wedge (1 \leq t_3 \leq n - 1)\}$ .

### A. Computation of Invalidate and Writeback Sets

**Invalidate Set:** Algorithm 1 shows how the invalidate set for a parallel task is computed. It is computed by forming the union of invalidate data sets corresponding to all statements within the parallel loop by iterating over each statement. For each statement  $S_i$ , first the source iterations of the dependence -  $I_{source}$  - whose target iterations are in the current slice for that statement -  $I_{current}^{S_i}$  - are determined by applying the inverse relation of the flow dependence. From this set, any of the source iterations that lie in the current slice -  $\bigcup_{S_j \in stmts} I_{current}^{S_j}$ , are removed from  $I_{source}$  because the source and target iterations are run on the same processor and no coherence instruction is needed. The array elements written by iterations of  $I_{source}$  are placed in the set of data elements for which invalidation coherence instructions must be issued to guarantee coherence. To this set is added the live-in list corresponding to data elements that come in live from outside the analyzed region.

---

#### Algorithm 1 Compute Invalidate Set

---

**Input:** Flow Dependences :  $\mathcal{D}_{flow}$ , Live-in read maps :  $\mathcal{D}_{live-in}$ , Current Iteration Slices:  $I_{current}$ , Write maps:  $r_{write}$   
**Output:** Statement and Invalidate set pairs:  $\mathcal{D}_{invalidate}^{S_i}$

- 1: **for all** statements -  $S_i$  **do**
- 2:  $\mathcal{D}_{invalidate}^{S_i} \leftarrow \emptyset$
- 3:  $I_{source} \leftarrow \mathcal{D}_{flow}^{-1}(I_{current}^{S_i}) \setminus (\bigcup_{S_j \in stmts} I_{current}^{S_j})$
- 4:  $\mathcal{D}_{inflow} \leftarrow \bigcup_{S_j \in stmts} r_{write}^{S_j}(I_{source})$
- 5:  $\mathcal{D}_{live-in\_data} \leftarrow \mathcal{D}_{live-in}(I_{current}^{S_i})$
- 6:  $\mathcal{D}_{invalidate}^{S_i} \leftarrow (\mathcal{D}_{inflow} \cup \mathcal{D}_{live-in\_data})$
- 7: **end for**

---

**Example:** The application of the algorithm to the running example results in the following invalidate set:  $\mathcal{D}_{invalidate}^{S_1} = \{[t_q, i_1] : 2 \leq i_1 \leq n\}$ . The array elements read in the parallel task are marked for invalidation.

**Writeback Set:** Algorithm 2 shows how we compute the writeback set for a parallel task that possibly has multiple statements in it. To find the writeback set corresponding to a statement  $S_i$ , first all target iterations ( $I_{target}$ ) of all dependences are identified whose source iterations lie in  $I_{current}^{S_i}$ . Those target iterations that are within the same parallel task -  $\bigcup_{S_j \in stmts} I_{current}^{S_j}$  are removed from  $I_{target}$  (line 3). Then the inverse dataflow relation is applied to this set and the intersection to the current iteration slice is computed (line 4) to identify the source iterations ( $I_{producer}$ ) in the slice that write values needed outside this slice. These values must be part of

the writeback set. Further, if a write by an iteration is the last write to a certain variable, it must also be written back since it represents a live-out value from the loop. The iterations that are not sources of any output dependencies produce live-out values. Such iterations are determined by forming the set difference between  $I_{current}^{Si}$  and domain of output dependences -  $dom \mathcal{D}_{output}$ .

---

### Algorithm 2 Compute Writeback Set

---

**Input:** Flow Dependences :  $\mathcal{D}_{flow}$ , Output Dependences :  $\mathcal{D}_{output}$ ,  
Current Iteration Slices:  $I_{current}$ , Write maps:  $r_{write}$

**Output:** Statement and Writeback set pairs:  $D_{writeback}^{Si}$

```

1: for all statements - Si do
2:    $D_{writeback}^{Si} \leftarrow \emptyset$ 
3:    $I_{target} \leftarrow \mathcal{D}_{flow}(I_{current}^{Si}) \setminus (\bigcup_{Sj \in stms} I_{current}^{Sj})$ 
4:    $I_{producer} \leftarrow \mathcal{D}_{flow}^{-1}(I_{target}) \cap I_{current}^{Si}$ 
5:    $D_{outflow} \leftarrow r_{write}^{Si}(I_{producer})$ 
6:    $I_{live\_out} \leftarrow I_{current}^{Si} \setminus dom \mathcal{D}_{output}$ 
7:    $D_{live\_out\_data} \leftarrow r_{write}^{Si}(I_{live\_out})$ 
8:    $D_{writeback}^{Si} \leftarrow (D_{outflow} \cup D_{live\_out\_data})$ 
9: end for

```

---

**Example:** The algorithm produces the following writeback set for the example in Fig. 2:  $D_{writeback}^{S1} = \{[t_q, i_1] : (t_p \leq tsteps - 2 \wedge 2 \leq i_1 \leq n - 1) \vee (t_p = tsteps - 1 \wedge 1 \leq i_1 \leq n - 1)\}$ .

For 0 to  $tsteps-2$  iterations of the outermost  $t1$  loop, only elements  $B[t2][2:n-1]$  need to be written back as they will be read in the next iteration of  $t1$  loop. Array cell  $B[t2][1]$  does not need to be written back because it is overwritten in a later  $t1$  iteration and its value is not read. But the very last write to  $B[t2][1]$ , i.e., when  $t1 = tsteps-1$  has to be written back as it is a live-out value of the loop.

a) *Code Generation:* The invalidate and writeback sets are translated to corresponding cache coherence instructions by generating a loop to traverse elements of the sets using a polyhedral code generator — ISL [26]. The invalidations and writebacks are combined into coherence range functions whenever elements of a set are contiguous in memory: when the inner-most dimension of the array is the fastest varying dimension of the loop.

#### B. Optimization: Analysis Cognizant of Iteration to Processor Mapping

The techniques described until now do not assume any particular mapping of iterations to processors. However, if a mapping of processors to iterations is known, many invalidations and write-backs could possibly be avoided. For example, in the code shown in Fig. 2, the flow dependence (mentioned in §II-C) is:  $S1[t_1, t_2, t_3] \mapsto S1[t_1 + 1, t_2, t_3 - 1]$ . If parallel iterations of the ‘ $t2$ ’ loop are mapped to processors such that an iteration with a particular ‘ $t2$ ’ value always gets mapped to the same processor, the source and target iterations of the flow dependence get executed on the same processor, making invalidations and write-backs due to the dependence unnecessary.

In order to incorporate this optimization, Algorithm 1 and 2 are modified to take iteration to processor mapping into

```

for (t1=0;t1<=tsteps-1;t1++)
#pragma omp parallel private (myid, t2, t3) {
  myid = omp_get_thread_num();
  for (t2=myid;t2<=n-1;t2+=8) {
    if (t1 == 0) {
      invalidate_range(&B[t2][2], sizeof(double)*(n-2));
    }
    invalidate_dword(&B[t2][n]);
    for (t3=1;t3<=n-1;t3++) {
      S1: B[t2][t3] = B[t2][t3+1] + 1;
    }
    if (t1 == tsteps-1) {
      writeback_range(&B[t2][1], sizeof(double)*(n-1));
    }
  }
}

```

---

Fig. 3: Optimized loop nest for SCC

account. Line 3 of Algorithm 1 is now changed to:

$$I_{source} \leftarrow \mathcal{D}_{flow}^{-1}(I_{current}^{Si}) \setminus (\bigcup_{Sj \in stms} I_{current}^{Sj} \cup I_{same\_proc})$$

and line 3 of Algorithm 2 is changed to:

$$I_{target} \leftarrow \mathcal{D}_{flow}(I_{current}^{Si}) \setminus (\bigcup_{Sj \in stms} I_{current}^{Sj} \cup I_{same\_proc}),$$

where  $I_{same\_proc}$  is the set of iterations that is executed on the same processor as the processor on which  $I_{current}$  is executed.

For the working example, let us say that the OpenMP scheduling clauses specify that iterations are cyclically mapped onto processors and the number of processors used is 8. Then, we encode that information into the following iteration to processor map:  $r_{i2p} = \{S1[t_1, t_2, t_3] \mapsto [t'_2] : t'_2 = t_2 \bmod 8\}$ . The parallel region code is all the iterations that are mapped to a parametric processor ‘myid’:  $I_{my\_proc} = r_{i2p}^{-1}(myid)$ . The iteration set  $I_{current}^{S1}$  is a subset of  $I_{my\_proc}$  with the values of the  $t1$  and  $t2$  loop iterators parameterized. Using the modified algorithms, the cache coherence code generated for the working example is presented in Fig. 3. In the optimized code, only the live-in data is invalidated: elements  $B[t2][2$  to  $n]$  at time-step  $t_1 = 0$ , only a single element –  $B[t2][n]$  at later time-steps, since other elements are written to by the same processor ensuring that the updated values are present in the processor’s private cache. Only the live-out data is written back at the last time-step:  $t_1 = tsteps - 1$ .

## IV. COMPILER OPTIMIZATION FOR IRREGULAR CODE

The irregular computations – code whose data flow and control flow may not be determined precisely at compile-time, are handled with a combination of compile-time and run-time techniques. We first describe a completely general scheme (§IV-A) that preserves data locality in caches within an *epoch* but not across *epochs*. Then, we present specialized methods (§IV-B, §IV-C) for certain classes of irregular code which preserve data locality across epoch boundaries.

### A. Bulk Coherence Operations: Basic Approach

The tasks that are executed in an epoch (interval between synchronization points) by construction do not have any dependences between them (otherwise, the dependences would induce serialization of tasks and hence, the tasks would have to be executed in different epochs). Therefore, all data accessed *within* an epoch can be safely cached and cache coherence is not violated.

For irregular applications that have non-affine references and hence, are not amenable to the analysis presented in

```

invalidate_all();
writeback_all();

```

Fig. 4: Coherence API for conservative handling

the previous section, software cache coherence is achieved conservatively: at the beginning of an epoch, the entire private cache is invalidated and at the end of the epoch, all data that are written in the current epoch (dirty words) are written to the shared cache. The coherence API functions shown in Fig. 4 are inserted in the parallel program at epoch boundaries to conservatively manage software coherence. The basic approach outlined above preserves intra-epoch cache data locality, but cannot exploit any temporal locality that exists across epoch boundaries.

### B. Inspector-Executors

Many scientific applications use sparse and irregular computations and are often iterative in nature and furthermore, the data access pattern remains the same across iterations. (Examples include programs for solving partial differential equations, irregular stencils, the conjugate gradient method for solving systems of linear equations which uses sparse matrix-vector multiplications, atmospheric simulations that use semi-regular grids).

```

while (converged == false) {
  #pragma omp parallel for
  for (i=0; i<n; i++) {
    read A[B[i]]; /*data-dependent access*/
  }

  #pragma omp parallel for
  for (i=0; i<n; i++) {
    write A[C[i]]; /*data-dependent access*/
  }
  /*Setting of converged variable not shown*/
}

```

Fig. 5: A time-iterated loop with irregular data references

For such code, we propose the use of *inspectors* to gather information on irregular data accesses so that coherence operations are applied only where necessary. The inspectors that are inserted in the parallel code are themselves parallel and are lock-free. The cost of inspectors is amortized by the ensuing selective invalidations of data and thus fewer unnecessary L1 cache misses over many iterations of the iterative computation. Fig. 5 shows an iterative code that has data-dependent references to a one-dimensional array, viz., A[B[i]] and A[C[i]]. We first illustrate the inspector approach for the simple example. The ideas are more generally applicable in the presence of multiple arrays and multi-dimensional arrays.

The inspector-code determines if a) the write performed at a thread has readers at other threads: if that is the case, the variable has to be written-back to shared cache so that other threads will be able to obtain the updated value of the variable. b) the variable being read at a thread was written by another thread: if yes, the variable has to be invalidated at the private cache so that the fresh value is retrieved from shared cache. Fig. 6 presents the inspector-inserted parallel

```

1  /*Inspector code begins*/
2  #pragma omp parallel for
3  for (i=0; i<n; i++) {
4  A_thread[i] = -1;
5  A_conflict[i] = 0;
6  writeback_word(&A_thread[i]);
7  writeback_word(&A_conflict[i]);
8  }
9  //Phase 1: Record writer thread ids
10 #pragma omp parallel for
11 for (i=0; i<n; i++) {
12   A_thread[C[i]] = myid;
13   writeback_word(&A_thread[C[i]]);
14 }
15 //Phase 2: Mark conflicted if
16 //writer and reader threads are not the same
17 #pragma omp parallel for
18 for (i=0; i<n; i++) {
19   invalidate_word(&A_thread[B[i]]);
20   if (A_thread[B[i]] != -1
21       && A_thread[B[i]] != myid) {
22     A_conflict[B[i]] = 1;
23     writeback_word(&A_conflict[B[i]]);
24   }
25 }
26 /*Inspector code ends*/
27
28 #pragma omp parallel
29 { invalidate_all(); }
30
31 while (converged == false) {
32   #pragma omp parallel for
33   for (i=0; i<n; i++) {
34     if (A_thread[B[i]] != -1
35         && A_thread[B[i]] != myid)
36       invalidate_word(&A[B[i]]);
37     read A[B[i]];
38   }
39
40   #pragma omp parallel for
41   for (i=0; i<n; i++) {
42     write A[C[i]];
43     if (A_conflict[C[i]] == 1)
44       writeback_word(&A[C[i]]);
45   }
46   /*Setting of converged variable not shown*/
47 }
48 #pragma omp parallel
49 { writeback_all(); }

```

Fig. 6: An iterative code with irregular data references for SCC system

code corresponding to the iterative code shown in Fig. 5 for execution on software managed caches.

Two shadow arrays — A\_thread and A\_conflict for array A that has data-dependent accesses are initialized (lines 4, 5). In the *first phase*, A\_thread records the ids of the threads that write to array cells (line 12). In the *second phase*, if an array cell is read by a thread different from the writer thread, the corresponding cell in A\_conflict array is set to 1 (line 22). Since the computation loops are parallel, the inspection is also carried out in parallel. Consequently, accesses to arrays A\_thread and A\_conflict are guarded with coherence instructions. If there are multiple readers for an array cell then more than one thread may set the respective cell of A\_conflict to 1 in phase two and multiple threads will write-back the same value, namely 1 to shared cache (in line 23). Since the same value is being written, any ordering of writes by different threads works.

Later in the computation loops, a thread invalidates a variable (line 36) before reading it if the variable has a writer (as opposed to read-only data) and that writer is a different thread. A thread after writing to a variable, writes it back (line

44) if the variable is marked conflicted.

### C. Exclusion of Read-Only Data from Coherence

```

/* Prologue begins */
writeback_all();
#pragma omp parallel
{ invalidate_all(); }
/* Prologue ends */
while (condition){
#pragma omp parallel
{
/* regular/irregular code */
}
}
/* Epilogue begins */
#pragma omp parallel
{ writeback_all(); }
invalidate_all();
/* Epilogue ends */

```

Fig. 7: A loop with bulk coherence operations at parallel region boundaries

For irregular code whose data access patterns potentially change with each iteration, we adopt a conservative approach which excludes read-only data from coherence enforcement and thus, is more accurate than a full invalidation and write-back approach outlined earlier. We consider *parallel regions* — parallel loops along with surrounding looping structures and perform analysis of the parallel region as a stand-alone unit. The *read-only data of the parallel region* need not be invalidated/written-back. Only those variables that are both written and read in the parallel region are invalidated and written-back at epoch boundaries.

For this scheme to work however, the following conditions have to be met:

- 1) None of the processors should have cached stale values of read-only data of the parallel region. (This could happen for example when, a program has a parallel region  $\mathcal{P}$  followed by a sequential segment  $Q$  and later a parallel region  $\mathcal{R}$ . And, variable  $x$  is read-only in  $\mathcal{P}$  and  $\mathcal{R}$ , but is modified in  $Q$ ).
- 2) Since, in the parallel region coherence is enforced only on data that are both read and written, for *written-but-not-read* data coherence operations should be introduced following the parallel region to ensure that future accesses to them get updated values.

To meet condition 1), a *prologue* is introduced that writes back all dirty words from the master thread and then does a full invalidation of caches at all threads. Condition 2) is fulfilled by writing-back all dirty words from all threads and doing a full-invalidation by the master thread in an *epilogue*. The code shown in Fig. 7 uses the outlined approach. Algorithm 3 presents the overall parallel-region analysis technique.

## V. OVERALL APPROACH

The overall compiler analysis operates by checking if it can apply optimizations in the following order (most constrained to unconstrained): 1) regular programs with static schedules (§III-B), 2) regular programs with dynamic schedules (§III-A), 3) inspector-executor (§IV-B): if the array-indexing variables

(e.g., in reference  $A[B[i]]$ , we refer to  $B[i]$  as the indexing variable) are not written inside the time-loop and the control flow, if any, does not contain variables that are written inside the time-loop. 4) exclusion of read-only data (§IV-C), finally, for the rest, 5) bulk coherence operations (§IV-A).

TABLE II: Simulator parameters

|                          |   |
|--------------------------|---|
| Processor chip           | 8-core multicore chip   |
| Issue width; ROB size    | 4-issue; 176 entries  |
| Private L1 cache         | 32KB Write-back, 4-way, 2 cycle hit latency                     |
| Shared L2 cache          | 1MB Write-back, 8-way, multi-banked<br>11 cycle round-trip time |
| Cache line size          | 32 bytes  |
| Cache coherence protocol | Snooping-based MESI protocol                                    |
| Main Memory              | 300 cycle round-trip time                                       |

## VI. EXPERIMENTAL EVALUATION

We evaluate the performance of compiler-generated coherence instructions for execution of parallel programs on software managed caches. The main goal of the compiler support developed in the paper is to insert coherence instructions — *invalidate* and *writeback* functions only where necessary. The conservative invalidations (of non-stale data) result in read misses which lead to degraded performance relative to a hardware coherence scheme. Therefore, to assess efficacy of the compiler techniques, we compare *read misses in L1 caches*, and *execution time* on software and hardware managed caches. (The number of misses at the shared cache is unaffected and will be the same for software and hardware cache coherence.)

Conservative coherence operations in software scheme increase accesses to the shared cache and also, cause increased traffic on the system bus. The hardware cache coherence protocol uses control messages to maintain coherence, which a software scheme does not. Therefore, if the software coherence mechanism results in comparable cache misses compared to a hardware protocol then, the software coherence also reduces network traffic and cache energy. We therefore measure the *number of words transferred on the system bus* and *cache energy* by software and hardware coherence systems.

### Algorithm 3 Generate Coherence Instructions using Parallel Region Analysis

---

**Input:** AST of Parallel region:  $\mathcal{P}$   
**Output:** AST of Parallel region for SCC:  $\mathcal{P}_{SCC}$

- 1: *Prologue*  $\leftarrow$  API to write-back all dirty words from master thread; API to invalidate entire cache of all threads
- 2: *Read\_Set*  $\leftarrow$  Arrays and scalars that are read in  $\mathcal{P}$
- 3: *Write\_Set*  $\leftarrow$  Arrays and scalars that are written in  $\mathcal{P}$
- 4: *Coherence\_Set*  $\leftarrow$   $Read\_Set \cap Write\_Set$
- 5: **for all** epoch code  $e \in \mathcal{P}$  **do**
- 6:   *Invalidate\_Set<sub>e</sub>*  $\leftarrow$   $Read\_Set_e \cap Coherence\_Set$
- 7:   *Writeback\_Set<sub>e</sub>*  $\leftarrow$   $Write\_Set_e \cap Coherence\_Set$
- 8:   Insert API for *Invalidate\_Set<sub>e</sub>* and *Writeback\_Set<sub>e</sub>*
- 9: **end for**
- 10: *Epilogue*  $\leftarrow$  API to write-back all dirty words from all threads; API to invalidate entire cache of master thread
- 11:  $\mathcal{P}_{SCC} \leftarrow$  Append {*Prologue*,  $\mathcal{P}$ , *Epilogue*}

---

TABLE I: Benchmarks. Legend: #PL: Number of Parallel Loops; #PLI: Number of Parallel Loops containing irregular accesses

| Benchmark  | Description                                  | Does application have irregular acc.? | #PL | #PLI | Techniques used                 |
|------------|--|---------------------------------------|-----|------|---------------------------------|
| gemm       | Matrix-multiply : $C = \alpha.A.B + \beta.C$ | No                                    | 2   | 0    | Polyhedral                      |
| gemver     | Vector Multiplication and Matrix Addition    | No                                    | 3   | 0    | Polyhedral                      |
| jacobi-1d  | 1-D Jacobi stencil computation               | No                                    | 2   | 0    | Polyhedral                      |
| jacobi-2d  | 2-D Jacobi stencil computation               | No                                    | 2   | 0    | Polyhedral                      |
| LU         | LU decomposition                             | No                                    | 1   | 0    | Polyhedral                      |
| trisolv    | Triangular solver                            | No                                    | 1   | 0    | Polyhedral                      |
| CG         | Conjugate Gradient method                    | Yes                                   | 3   | 1    | Inspector-Executor + Polyhedral |
| backprop   | Pattern recognition using unstructured grid  | Yes                                   | 2   | 0    | Bulk + Polyhedral               |
| hotspot    | Thermal simulation using structured grid     | Yes                                   | 2   | 0    | Bulk + Polyhedral               |
| kmeans     | Clustering algorithm used in data-mining     | Yes                                   | 1   | 1    | Exclusion of RO data            |
| pathfinder | Dynamic Programming for grid traversal       | Yes                                   | 1   | 1    | Exclusion of RO data            |
| srad       | Image Processing using structured grid       | Yes                                   | 2   | 2    | Inspector-Executor              |

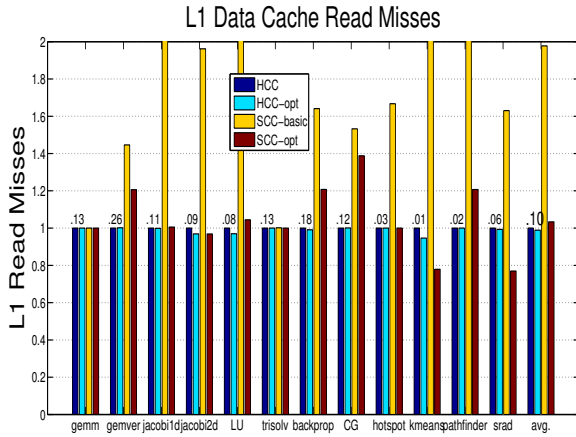


Fig. 8: L1 data cache read misses (lower, the better). The L1 read miss ratios for HCC are also shown.

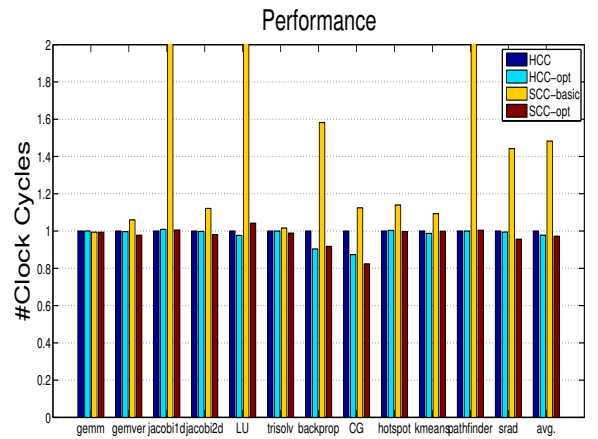


Fig. 9: Execution time (lower, the better)

### A. Benchmarks

The benchmark programs used for the experiments and their characteristics are listed in Table I. The benchmark programs — gemm, gemver, jacobi-1d, jacobi-2d, LU, trisolv are taken from PolyBench benchmark suite [25]. The PolyBench benchmark suite is a collection of widely used linear algebra, and stencil code. The code are parallelized using a polyhedral compiler – PoCC [24]. All array references in the PolyBench programs are affine and coherence instructions are generated using Polyhedral techniques presented in Section III.

The backprop, hotspot, kmeans, pathfinder, srad applications are taken from Rodinia suite [4], and they contain affine as well as irregular data references. The Rodinia suite provides parallel programs from various application domains. At the parallel region boundaries in Rodinia applications, bulk coherence instructions (`invalidate_all` and `writeback_all`) are applied (Section IV-A) while the parallel regions are optimized. Inspector-executor method (Section IV-B) is used for irregular data references in CG and srad applications. Exclusion of read-only data optimization described in Section

IV-C is employed in kmeans and pathfinder code. The parallel loops in backprop and hotspot benchmarks are amenable to polyhedral analysis and therefore, bulk coherence operations are inserted at the beginning and end of parallel regions, and coherence operations in parallel loops are derived using polyhedral algorithms (Section III).

### B. Set-up

The snooping-bus MESI protocol hardware coherence (referred to as *HCC* in the following text), and software cache coherence (referred to as *SCC*) have been implemented in an architectural multi-processor simulator — SESC [23]. Details of the simulator setup are described in Table II.

We compare performance and energy of the following four coherence schemes:

- 1) **HCC**: Parallel programs are executed using MESI hardware coherence.
- 2) **SCC-basic**: The coherence instructions are inserted without iteration-to-processor aware analysis for affine references and without the use of inspector-executor or read-only data exclusion scheme for irregular accesses. That

is, coherence instructions are generated with methods described in Sections III-A and IV-A only without further optimizations. The resulting code are run on software managed caches.

- 3) **SCC-opt:** The coherence management is optimized using compiler optimizations presented, and the resulting programs are executed on software managed caches.
- 4) **HCC-opt:** To study if any optimizations applied to SCC code (such as explicit mapping of iterations to processors) can also benefit the benchmarks for hardware coherence, SCC-opt programs are adapted to run on HCC systems: coherence operations and any *inspectors* inserted are removed from SCC-opt code and these variants are run on the HCC system.

The performances of only parallel parts of benchmarks are measured — sequential initialization and finalization code are excluded from measurements because the performance of sequential code is expected to be the same on SCC and HCC systems. Threads are pinned to cores for both schemes.

### C. Performance Results

Fig. 8 plots the read misses in L1 cache; Fig. 9 shows the execution time. The number of L1 read misses and execution cycles are normalized with respect to HCC statistics (the number of misses and execution cycles of HCC is considered 1). The L1 read miss ratios (fraction of L1 reads that are misses) for HCC are also indicated in the graph. On average (geometric mean) across benchmarks, HCC-opt has the same number of cache misses as HCC; SCC-basic suffers 98% more misses and SCC-opt experiences only a 3% increase (avg. column in the graph). The geometric mean of normalized execution time for the three variants — HCC-opt, SCC-basic, and SCC-opt are, 0.97, 1.48, and 0.97 respectively. We observe that SCC-opt greatly improves performance over SCC-basic and brings down cache misses comparable to those of HCC. Further, performance of HCC-opt is very similar to that of HCC.

The *gemm* and *trisolv* benchmarks exhibit the so-called *communication free* parallelism: the outer loop in these code is parallel. Therefore, there is no communication between processors induced by data dependences. All code variants of *gemm* and *trisolv* have virtually the same number of cache misses and execution cycles. In applications that have irregular references, namely *backprop*, *CG*, *hotspot*, *kmeans*, *pathfinder*, *srad*, the parallel region boundaries are guarded with full-invalidation and full-writeback instructions (described in IV-C) The affine accesses in the parallel regions are optimized; irregular accesses are handled using *inspectors* or invalidation and write-back of entire arrays that are both written and read in the parallel region (read-only arrays and scalars are excluded). For *backprop* and *pathfinder*, full invalidation of cache at parallel region boundaries results in some loss of data locality which results in increased L1 cache read misses. The *CG* and *srad* benchmarks have iterative loops and irregular accesses whose indexing structures do not change across iterations. Therefore, for those two benchmarks, *inspector* code are inserted for

deriving coherence operations. The inspectors contribute to a certain number of L1 read misses. The reduced cache misses in SCC-opt of *srad* compared to its HCC counterpart is an artifact of the interaction that exists between cache coherence and cache replacement policy (LRU): false-sharing in HCC can cause soon-to-be-reused data to be evicted, which favors SCC. The migratory writes may sometimes cause invalidations of not-to-be-reused data and thus, making way for other to-be-reused data and this benefits HCC. Conversely, the *gemv* is an example of hardware cache coherence working to HCC advantage, where the number of misses for HCC is lower compared to SCC-opt.

The running time (depicted in Fig. 9) shows a strong correlation between L1 cache read misses and performance. In HCC, the snooping overhead plays a significant role in determining execution time: In our implementation, we assign 1 cycle to a read/write snooping request. In SCC, each coherence instruction incurs a two-cycle overhead. In addition to these overheads, there may be additional overheads depending upon the response to a snooping request in HCC (e.g., a read request may return an updated value from another processor) and the number of cache lines specified in the coherence instruction in SCC — each cache line incurs a 2-cycle delay. Because of removal of hardware cache coherence, we observe a 3% performance gain for SCC-opt over HCC on average.

**Discussion:** The performance results obtained for HCC and SCC schemes are sensitive to architectural choices made in the simulator implementation. And, we have opted for architectural choices that favor HCC even though on a real system they may be impractical or too costly. E.g., we have allotted 1-cycle delay for a snooping request and on a real system it might take multiple cycles. The implemented HCC protocol in the simulator concurrently sends a snoop request to other cores, and also a memory request to L2 cache. Alternatively, the L1 cache can also be designed to send a memory request to L2 cache after a snoop miss, but this will increase the delay when there is a snoop miss.

### D. Energy Results

**Bus data transfers:** Fig. 10 shows the traffic (number of words transferred) on the system bus for different schemes. All values are normalized with respect to HCC. The average number of words transferred per cycle (obtained by dividing total number of words with number of execution cycles) for HCC is also shown. For hardware coherence scheme, the traffic on the bus includes snoopy-bus coherence related exchange of messages, transfers between private L1 caches and shared L2 cache triggered by cache misses at L1 and replacement of cache lines at L1. For SCC, this includes data transfers between L1 caches and L2 cache prompted either by L1 misses and evictions, or invalidation and writeback coherence instructions. The HCC normalized data transfers on the bus for HCC-opt, SCC-basic, and SCC-opt are 0.99, 1.46, and 0.99 respectively, on average (geo-mean). In *backprop* and *srad*, SCC-opt does fewer write-backs to L2 cache compared to HCC; the L1 cache misses are lower for SCC-opt in the case



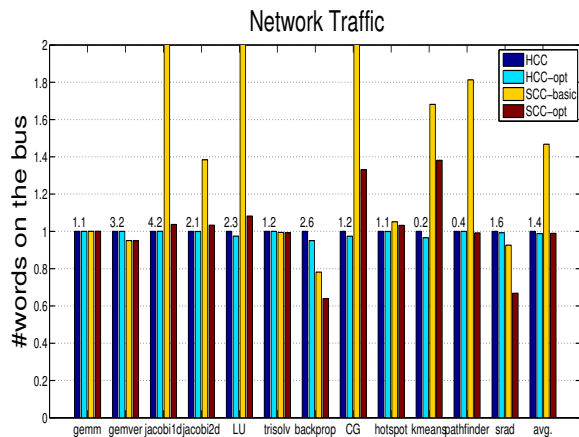


Fig. 10: Traffic on the system bus (lower, the better). Average number of words per cycle for HCC is also shown.

of *srad*. Consequently, SCC-opt incurs fewer data transfers in *backprop* and *srad*. Conservative writebacks in *kmeans* increases the traffic on the bus for SCC-opt compared to HCC.

**L1 and L2 cache energy:** The cache SRAM is a major consumer of energy in a processor. We compare cache energy consumption for HCC and SCC-opt schemes based on the number of accesses to tag SRAMs and data SRAMs. Using the SESC simulator, event counts for all relevant activities in the L1 and L2 caches are collected to account for all tag and data accesses to SRAMs. CACTI [27] is used to obtain the energy per access for tag and data for each cache level. The L1 cache employs dual ported SRAM to service snoop requests quickly. For SCC also we used the same dual ported SRAM for a fair comparison (per-access cost is a function of, inter alia, number of ports). The L1 cache accesses tag and data together for local processor requests while for snooping requests it accesses data SRAM only after tag hit. The L2 cache is configured to be in sequential access mode — it starts to access data SRAM after tag matching. We did not consider main memory energy because main memory accesses would be the same for both HCC and SCC schemes.

Fig. 11 plots relative energy consumption in caches for hardware and software cache coherence approaches: energy expenditure by HCC is considered 1 and energy dissipation by SCC-opt is scaled with respect to HCC. The break-down of energy expended in L1 and L2 caches is indicated. On average (arithmetic mean) SCC-opt energy consumption in caches is 5% less than that of HCC. Most of the savings in SCC-opt come from two sources: elimination of snooping requests in L1 cache, and reduction in the number of *writeback words* by partial line transfers (only dirty words are written back to shared L2 cache in a software managed cache as opposed to entire cache lines which are the granularity of communication for HCC). We also observe that energy spent in all L1 caches together is around 86%, while the rest — 14% is expended in L2 cache.

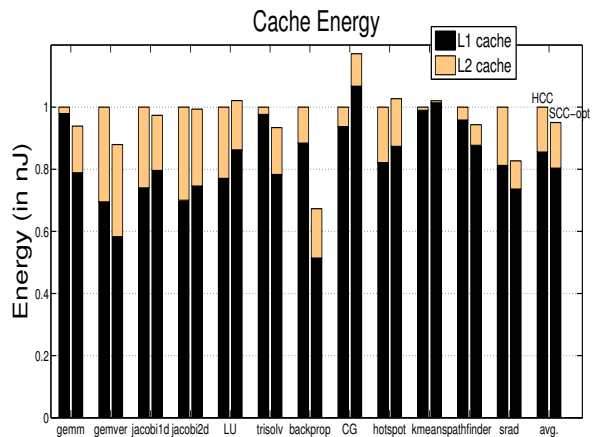


Fig. 11: L1 and L2 Cache Energy (lower, the better). The first bar shows HCC energy and second bar SCC-opt energy

## VII. RELATED WORK

Some prior studies [7], [5], [8] have developed compiler analysis techniques to generate cache coherence instructions for software managed caches. The work in this paper distinguishes itself from prior efforts both by being more general as well as more precise, as we elaborate below.

Cheong et al. [5] use data flow analysis to classify every reference to shared memory either as a *memory-read* or a *cache-read*. A read reference is marked as a memory-read if the compiler determines that the cache resident copy of the data might have become stale, otherwise the reference is marked as cache-read. A limitation of that work is that the data flow analysis is carried out at the granularity of arrays, which will result in invalidations for an entire array even if two processors are accessing distinct parts of it. Choi et al. [7] propose to improve inter-task locality in software managed caches by using additional hardware support: the current epoch number is maintained at runtime using an *epoch counter* and each cache word is associated with a *time-tag* which records the epoch number in which the cache copy is created. Then they develop the so-called epoch flow graph technique to establish conditions under which it can be guaranteed that the cached copy of a variable is not stale. The analysis here too treats an entire array as a single variable. Lim et al [20] build on this work and combine data prefetching and cache coherence with a customized data prefetching hardware support.

Darnell and his colleagues [8] perform array subscript analysis to gather more accurate data dependence information and then aggregate cache coherence operations on a number of array elements to form *vector* operations. The method however can handle only simple array subscripts: only loop iterators are allowed as subscripts.

O’Boyle et al. [22] develop techniques to identify data for Distributed Invalidation (DI) for *affine loops* that are run on distributed shared memory architectures. The DI scheme uses

a directory to maintain coherence, and where possible it seeks to eliminate invalidation messages from the directory to remote copies and associated acknowledgments. Their analysis to minimize invalidation messages has similarities to our analysis for minimizing invalidations. But the coherence equations in DI place some restrictions on the kinds of affine loops that can be analyzed: for example, conditional execution within the loop is disallowed, and increments of loop iterators must be unity. The approach presented in this paper efficiently handles arbitrary affine loops including those whose iterator values are lexicographically decreasing, that have a non-unit increment-count, or have a modulus operator etc., and conditionals are permitted. The DI work does not involve writebacks, which however are a part of our software cache coherence model, and we develop techniques to optimize writebacks as well. We also optimize irregular code using a number of techniques including an inspector-executor approach, while such code are not optimized in the DI scheme.

Inspector-executor approaches have been used in the context of parallelization (e.g., [9]), run-time reorderings [10] but to our knowledge have not previously been developed for optimizing for cache coherence.

There also have been numerous works exploring software cache coherence and, simplifying hardware cache coherence schemes by seeking assistance from software. We discuss a few of those efforts here and note that none of them develop compiler algorithms to instrument the needed software instructions automatically. Instead, they rely on the programmer or complementary compiler support such as the one proposed in this paper to achieve the required program modifications.

Kontothanassis et al. [18] present a software cache coherence protocol with page granularity in large scale machines. The compiler analysis developed in this paper supports fine-grained coherence at the level of variables. DeNovo [6] simplifies complicated hardware cache coherence protocols by enforcing a disciplined parallel programming model. The compiler support proposed in this work can complement the DeNovo project in automatically identifying self-invalidation regions. Kim et. al [17] provide an architectural design for incoherent cache hierarchies and propose two programming approaches – inter-block and intra-block. The compiler work in this paper is complementary to their architecture-centric work.

## VIII. CONCLUSION

The complexity of developing efficient hardware coherence protocols for emerging manycore heterogeneous systems makes software controlled coherence schemes attractive. However, a significant challenge for software controlled cache coherence is that of generation of efficient coherence instructions. The automatic coherence management and optimization approaches developed in the paper advance compiler technology towards making software cache coherence a viable solution on shared-memory multiprocessor systems. Simulation results demonstrate the effectiveness of the compiler algorithms

in achieving performance and cache-energy comparable to that of a hardware cache coherence scheme.

## REFERENCES

- [1] D. Abts, S. Scott, and D. J. Lilja. So many states, so little time: Verifying memory coherence in the cray x1. In *IPDPS*, 2003.
- [2] J. Archibald and J.-L. Baer. Cache coherence protocols: Evaluation using a multiprocessor simulation model. *ACM Transactions on Computer Systems (TOCS)*, 4(4), 1986.
- [3] N. Carter, A. Agrawal, S. Borkar, R. Cledat, H. David, D. Dunning, J. Fryman, I. Ganev, R. Golliver, R. Knauerhase, R. Lethin, B. Meister, A. Mishra, W. Pinfold, J. Teller, J. Torrellas, N. Vasilache, G. Venkatesh, and J. Xu. Runnemed: An architecture for ubiquitous high-performance computing. In *HPCA*, 2013.
- [4] S. Che, M. Boyer, J. Meng, D. Tarjan, J. Sheaffer, S.-H. Lee, and K. Skadron. Rodinia: A benchmark suite for heterogeneous computing. In *IEEE International Symposium on Workload Characterization*, 2009.
- [5] H. Cheong and A. V. Vaidenbaum. A cache coherence scheme with fast selective invalidation. *ISCA*, 1988.
- [6] B. Choi, R. Komuravelli, H. Sung, R. Smolinski, N. Honarmand, S. V. Adve, V. S. Adve, N. P. Carter, and C.-T. Chou. Denovo: Rethinking the memory hierarchy for disciplined parallelism. *PACT*, 2011.
- [7] L. Choi and P.-C. Yew. A compiler-directed cache coherence scheme with improved intertask locality. In *Supercomputing*, 1994.
- [8] E. Darnell, J. M. Mellor-Crummey, and K. Kennedy. Automatic software cache coherence through vectorization. In *International Conference on Supercomputing*, 1992.
- [9] R. Das, P. Havlak, J. Saltz, and K. Kennedy. Index array flattening through program transformation. In *SC*, 1995.
- [10] C. Ding and K. Kennedy. Improving cache performance in dynamic applications through data and computation reorganization at run time. *PLDI*, 1999.
- [11] P. Feautrier. Dataflow analysis of array and scalar references. *IJPP*, 20(1), 1991.
- [12] P. Feautrier. Some efficient solutions to the affine scheduling problem: I. one-dimensional time. *IJPP*, 21(5):313–348, 1992.
- [13] M. Ferdman, P. Lotfi-Kamran, K. Balet, and B. Falsafi. Cuckoo directory: A scalable directory for many-core systems. In *HPCA*, 2011.
- [14] S. Kaxiras and G. Keramidas. SARC coherence: Scaling directory cache coherence in performance and power. *Micro, IEEE*, 30(5), 2010.
- [15] S. Kaxiras and A. Ros. A new perspective for efficient virtual-cache coherence. *ISCA '13*, 2013.
- [16] J. H. Kelm, D. R. Johnson, W. Tuohy, S. S. Lumetta, and S. J. Patel. Cohesion: A hybrid memory model for accelerators. *ISCA*, 2010.
- [17] W. Kim, S. Tavarageri, P. Sadayappan, and J. Torrellas. Architecting and programming a hardware-incoherent multiprocessor cache hierarchy. *IPDPS*, 2016.
- [18] L. Kontothanassis and M. Scott. Software cache coherence for large scale multiprocessors. In *HPCA*, 1995.
- [19] J. Laudon and D. Lenoski. The sgi origin: a ccnuma highly scalable server. In *ACM SIGARCH Computer Architecture News*, volume 25, 1997.
- [20] H.-B. Lim and P.-C. Yew. Efficient integration of compiler-directed cache coherence and data prefetching. In *IPDPS*, 2000.
- [21] T. G. Mattson and et.al. The 48-core scc processor: The programmer's view. *SC*, 2010.
- [22] M. O'Boyle, R. Ford, and E. Stohr. Towards general and exact distributed invalidation. *Journal of Parallel and Distributed Computing*, 63(11), 2003.
- [23] P. M. Ortega and P. Sack. Sesc: Superescalar simulator. In *Euro micro conference on real time systems*, 2004.
- [24] PoCC: the Polyhedral Compiler Collection. <http://sourceforge.net/projects/pocc/>.
- [25] PolyBench: The Polyhedral Benchmark suite. <http://sourceforge.net/projects/polybench/>.
- [26] S. Verdoolaege. isl: An integer set library for the polyhedral model. *Mathematical Software–ICMS 2010*, pages 299–302, 2010.
- [27] S. J. E. Wilton and N. P. Jouppi. Cacti: An enhanced cache access and cycle time model. *IEEE Journal of Solid-State Circuits*, 31, 1996.
- [28] X. Zhou, H. Chen, S. Luo, Y. Gao, S. Yan, W. Liu, B. Lewis, and B. Saha. A case for software managed coherence in manycore processors. In *2nd USENIX Workshop on Hot Topics in Parallelism HotPar10*, 2010.