CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2007; **19**:1481–1496
Published online 24 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe.1128

Automatic analysis of inefficiency patterns in parallel applications

Felix Wolf[1,*,†], Bernd Mohr[1], Jack Dongarra[2] and Shirley Moore[2]

[1] *Forschungszentrum Jülich, ZAM, 52425 Jülich, Germany*
[2] *University of Tennessee, ICL, 1122 Volunteer Boulevard, Suite 413, Knoxville, TN 37996-3450, U.S.A.*

SUMMARY

Event tracing is a powerful method for analyzing the performance behavior of parallel applications. Because event traces record the temporal and spatial relationships between individual runtime events, they allow application developers to analyze dependences of performance phenomena across concurrent control flows. However, in view of the large amounts of data generated on contemporary parallel machines, the depth and coverage of a purely manual analysis is often limited. Our approach automatically searches event traces for patterns of inefficient behavior, classifies detected instances by category, and quantifies the associated performance penalty. This enables developers to study the performance of their applications at a high level of abstraction, while requiring significantly less time and expertise than a manual analysis. Copyright © 2006 John Wiley & Sons, Ltd.

Received 19 April 2006; Accepted 28 July 2006

KEY WORDS: performance tools; event tracing; pattern search

1. INTRODUCTION

High-performance computing is playing an increasingly critical role in advanced scientific research as simulation and computation are becoming widely used to augment and/or replace physical experiments.

*Correspondence to: Felix Wolf, Forschungszentrum Jülich, ZAM, 52425 Jülich, Germany.
†E-mail: f.wolf@fz-juelich.de

Contract/grant sponsor: DOE SciDAC Program; contract/grant number: DE-FC02-01ER25490
Contract/grant sponsor: DOE MICS Program; contract/grant number: DE-FG02-01ER25510
Contract/grant sponsor: DoD HPCMP PET Program; contract/grant number: N62306-01-7110
Contract/grant sponsor: European IST Working Group APART; contract/grant number: IST-2000-28077
Contract/grant sponsor: Helmholtz Research Programme Scientific Computing—Key Technologies

However, the gap between peak and achieved performance for scientific applications running on parallel systems has grown considerably in recent years. The complex architecture of parallel systems and the interdependence between different components and layers in conjunction with communication structures imposed by the algorithm, present difficult challenges for the performance optimization of scientific applications. Tools are needed that collect and present relevant information on application performance at a high level of abstraction so as to enable developers to easily identify and determine the causes of performance bottlenecks.

Event tracing is a powerful method for analyzing the performance behavior of parallel applications. Time-stamped events, such as entering a function or sending a message, are recorded at runtime and analyzed afterward with the help of software tools. Graphical trace browsers, such as VAMPIR [1] and Paraver [2], allow the fine-grained investigation of parallel performance behavior using a zoomable time-line display, as well as providing statistical summaries of communication behavior. Because event traces record the temporal and spatial relationships between individual runtime events, they allow application developers to analyze dependences of performance phenomena across concurrent control flows.

However, in view of the large amounts of data generated on contemporary parallel machines, the depth and coverage of the visual analysis offered by a browser is limited as soon as it targets more complex patterns not included in the statistics generated by such tools. In this paper, we present an alternative to manually scanning the time-line display. Our approach automatically searches event traces of MPI and OpenMP programs for patterns of inefficient behavior, classifies detected instances by category, and quantifies the associated performance penalty. This allows developers to study the performance of their applications at a high level of abstraction, while requiring significantly less time and expertise than a manual analysis. The abstraction level is achieved through specifying complex patterns that embody higher-level behavior related to the parallel programming model. Moreover, using wavefront algorithms as an example, we show that the semantic content of these patterns can be further increased by correlating their occurrences with performance-critical phases of the parallelization strategy used in an application. Such phases manifest themselves as another class of patterns in the event trace and can be recognized using knowledge of virtual adjacency relationships between individual processes.

Our approach constitutes the core of the KOJAK toolkit [3] and is primarily implemented in the EXPERT trace analyzer component [4]. EXPERT maps the execution-time penalty caused by each pattern onto a three-dimensional space consisting of the following hierarchical dimensions: (i) performance property (i.e. a more general term used instead of 'performance problem' that also includes non-negative performance aspects such as computation), (ii) call path, and (iii) system resource (e.g. a process). A graphical browser allows a convenient in-depth study of this space at varying levels of granularity.

The remainder of this document is organized as follows. After considering related work in Section 2, we give a brief overview of the KOJAK toolkit and explain the basic process of analyzing a trace file in Section 3. Then, in Section 4, we illustrate how patterns of inefficient execution can be specified and study the underlying abstraction mechanisms along with a realistic example. The benefits of using topological knowledge to explain the occurrence of inefficiency patterns in terms of the parallelization strategy of a program are described in Section 5. Finally, we present a conclusion followed by an outlook on future work in Section 6.

2. RELATED WORK

This work is embedded in the ESPRIT/IST working group APART. One of the core results of APART is the APART Specification Language (ASL) [5], which provides a formal notation to describe performance properties of parallel applications. A performance property represents a not necessarily negative aspect of an application's performance, such as synchronization or computation. In ASL, performance properties are specified as conditions referencing performance-related data, such as source-code entities or certain measurements, in a uniform way by means of an object-oriented data model. A severity expression quantifies a property's impact on the overall performance, while a confidence value quantifies the condition's reliability. The notion of a performance property strongly influenced EXPERT's modular architecture, which is based on encapsulating each performance property for which the search process is looking in a separate C++ class, offering methods to control its evaluation. Conversely, the EXPERT approach inspired mechanisms in ASL to define performance properties based on trace data.

Also stimulated by ASL, Fahringer and Seragiotto Júnior [6] designed a language called JavaPSL to specify performance properties in the Aksum tool based on the Java programming language. In contrast to EXPERT, which concentrates on compound-event analysis, JavaPSL puts emphasis on the definition of performance properties based on existing properties (i.e. by defining metaproperties) using advanced concepts of the Java language, such as polymorphism, abstract classes, and reflection.

Fürlinger *et al.* [7] created another ASL-inspired tool called Periscope that conducts on-line performance analysis based on a hierarchical network of agents transforming lower-level information stepwise into higher-level information.

KappaPI 2 by Jorba *et al.* [8] searches trace files of message-passing applications for patterns very similar to those used in our approach. A distinctive feature of KappaPI is that it generates recommendations on how to improve the performance using knowledge of bottleneck use cases in combination with source-code analysis.

Vetter [9] automatically identified wait states in MPI point-to-point communication based on machine learning techniques. He traced individual message-passing operations and then classified each individual communication event using a decision tree. The decision tree has been previously trained by microbenchmarks that demonstrate both efficient as well as inefficient performance behavior. As opposed to this approach, EXPERT draws conclusions from the temporal relationships of individual events in a platform-independent way that does not require any training prior to analysis.

An alternative approach to describing complex event patterns was devised by Bates. The proposed Event Definition Language (EDL) [10] focuses on specifying incorrect behavior of distributed systems. It allows compound events to be defined in a declarative manner based on extended regular expressions, where primitive events are clustered to higher-level events using certain formation operators. However, EDL's suitability for compound events that are associated with some kind of state, such as those targeted by EXPERT, is limited.

The multidimensional hierarchical decomposition of the search space for performance problems has a long tradition. Miller *et al.* [11] developed the W^3 Search Model as the basis for the on-line performance-analysis performed by Paradyn. The W^3 model describes performance behavior in a space spawned by the dimensions of performance problem, program resource, which may also include the call graph, and time. Performance problems are expressed in terms of a threshold and one or more metrics such as CPU time, blocking time, message rates, I/O rates, or number of active processors.

The different metrics can be specified in a flexible manner using the metric-description language (MDL). The main accomplishment of EXPERT in contrast to Paradyn is the description of performance problems in terms of complex event patterns that go beyond counter-based metrics. Also, the uniform mapping of arbitrary performance behavior onto fractions of the overall execution time allows the correlation of different behavior in a single view.

Topological knowledge has been used to highlight certain aspects of parallel performance. For example, Ahn and Vetter [12] mapped counter data onto the virtual topology of the SWEEP3D benchmark to identify clusters of related behavior by statistical means. Also, topological knowledge has proven to be beneficial for semantic debugging of parallel applications. Huband and McDonald describe a trace-based debugger called DEPICT that exploits topological information to identify processes with logically similar behavior in traces of MPI applications and to display semantic differences among these groups [13]. The comparison is based on the order and number of events. Furthermore, DEPICT's ability to automatically identify the virtual topology using graph-distance measures is also of interest to our work.

3. THE KOJAK TOOLKIT

KOJAK is an automatic performance evaluation system for MPI, OpenMP, SHMEM, and hybrid applications written in C/C++ or Fortran. KOJAK generates event traces from running applications and automatically searches them off-line for execution patterns indicating inefficient performance behavior. KOJAK is jointly developed by Forschungszentrum Jülich, Germany, and the University of Tennessee, U.S.A. Figure 1 gives an overview of KOJAK's architecture, its components, and the overall process of analyzing a trace file. The process involves three major parts: (i) a semi-automatic multi-level instrumentation of the user application, (ii) the execution on the target platform, and (iii) the automatic analysis of the generated trace file.

The component mainly responsible for trace generation is the EPILOG runtime system. Event traces generated by EPILOG capture: MPI point-to-point, collective, and one-sided communication; OpenMP parallelism change, parallel constructs, and synchronization; and SHMEM one-sided and collective communication. In addition, data from hardware counters accessed using the PAPI library [14] can be recorded as part of the event traces. To make measurements with the EPILOG system, the application must be instrumented at specific points to activate EPILOG library calls. These points usually include the entries and exits of various code regions, such as functions, OpenMP constructs, and MPI calls.

The automatic instrumentation of the user code is supported in three different ways, depending on the availability of certain compilers and third-party tools: (i) using a compiler-supplied profiling interface, (ii) using TAU [15], or (iii) using DPCL [16]. In addition, the user is free to instrument arbitrary user regions manually by placing POMP directives after the entry point and before all exit points. The POMP directives are later processed by OPARI [17], which is also responsible for the automatic instrumentation of OpenMP constructs. MPI functions are instrumented fully automatically by interposing a wrapper library.

During execution, the instrumented code generates several trace files, one for each process or thread. Execution time overheads ranging between about 1 and 10% have been reported in Wolf [18]. After termination, the local traces are merged into a single global trace file. In the absence of a global clock, time stamps are synchronized off-line using a linear interpolation between offset measurements

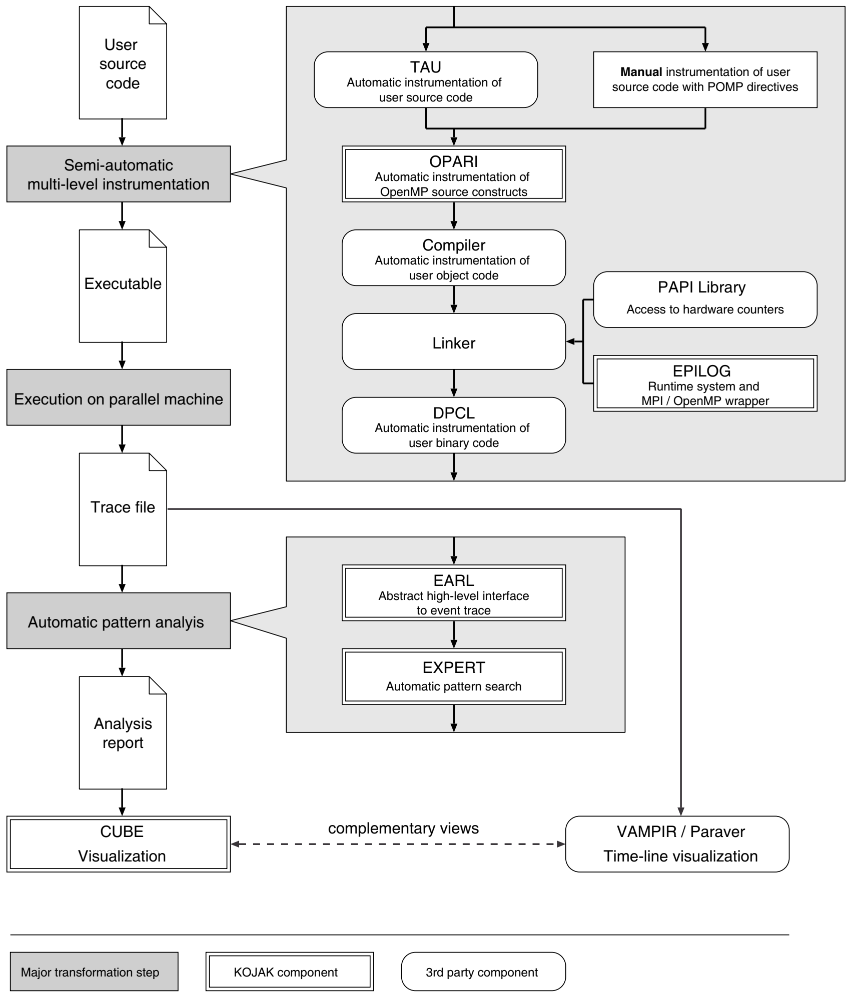

User source code
Semi-automatic multi-level instrumentation
Executable
Execution on parallel machine
Trace file
Automatic pattern analysis
Analysis report
CUBE Visualization
TAU Automatic instrumentation of user source code
Manual instrumentation of user source code with POMP directives
OPARI Automatic instrumentation of OpenMP source constructs
Compiler Automatic instrumentation of user object code
PAPI Library Access to hardware counters
Linker
EPILOG Runtime system and MPI / OpenMP wrapper
DPCL Automatic instrumentation of user binary code
EARL Abstract high-level interface to event trace
EXPERT Automatic pattern search
complementary views
VAMPIR / Paraver Time-line visualization
Major transformation step
KOJAK component
3rd party component

Figure 1. KOJAK architecture.

taken during the program start and program termination, maintaining the correct logical event order to a certain degree. If this is not sufficient, the correct causal order can be reestablished based on logical clocks [19].

After the postprocessing has been completed, the global trace is subjected to an off-line analysis performed by KOJAK's EXPERT component, which attempts to identify specific performance properties. Internally, EXPERT represents performance properties in the form of execution patterns that model inefficient behavior. These patterns are used during the analysis process to recognize, classify, and quantify inefficient behavior in the application. The performance properties addressed by EXPERT include inefficient use of the parallel programming models MPI and OpenMP as well as low CPU and memory performance. The analysis process automatically transforms the traces into a compact call-path profile that includes the execution time penalties caused by the different patterns. Section 4 covers the pattern analysis in more detail.

The call-path profile can be viewed using the CUBE performance browser (Figure 5). CUBE is a generic tool for displaying a multidimensional performance space consisting of the following dimensions: (i) performance property, (ii) call path, and (iii) system resource. Each dimension is represented as a tree browser that can be collapsed or expanded to achieve the desired level of granularity or specialization. The tree browsers are coupled such that the penalty caused by a particular performance property can be broken down by call path and process or thread. An algebra utility [20] supports the comparison of analysis results between different experiments (e.g. to verify optimizations).

In addition, the automatic analysis can be combined with manual time-line analysis using VAMPIR or Paraver to investigate the context of the previously identified patterns in more detail. For this purpose, KOJAK includes appropriate trace-format conversion utilities.

4. PATTERN ANALYSIS

The central idea behind the KOJAK approach is to identify performance properties by searching event traces recorded during execution for patterns of inefficient behavior. This permits the use of the following techniques: (i) classifying the behavior that leads to a performance degradation, and (ii) quantifying its impact on the overall performance. The particular way EXPERT specifies these patterns internally enables us to capture very complex situations not covered by the previously mentioned trace-visualization tools or by typical profiling tools.

EXPERT specifies patterns as *compound events*. A compound event is a set of events appearing in the trace file that satisfy conditions related to a specific performance problem. These conditions are expressed in terms of an event model suitable for describing the execution of a parallel program. The execution of a program, as represented by an event trace, is modeled as a chronologically sorted sequence of events representing actions relevant to the purpose of the observation. Actions of interest are: the sending and receiving of point-to-point messages; entering and exiting different types of code regions, such as user functions, OpenMP constructs, and MPI (collective) operations; and synchronization operations, such as acquiring and releasing OpenMP locks. Each action is represented by a different event type. The basic structure of the event trace along with the applied event-type system is called the *basic event model*.

4.1. Abstraction mechanisms

As the compound events targeted by our analysis often involve complex inter-event relationships referring to certain aspects of the execution state, such as message queues between different locations or call stacks for single locations, the basic event model is not convenient to describe the corresponding patterns as the only relationship explicitly provided at this level is the global temporal order within the accuracy bounds of the time synchronization.

To be better able to express complex relationships among the constituents of a compound event, the basic event model has been enhanced by adding higher-level abstractions defined on top of the basic event model. The abstractions offered by the *enhanced event model* are implemented in the EARL trace access layer [21], which is a class library used by EXPERT during the pattern search to access the trace file (Figure 1, above EXPERT). The main purpose of EARL is to simplify the specification of execution patterns representing performance properties within the EXPERT analyzer and, thus, to allow the easy extension and customization of the pattern base used in the analysis process. As opposed to a raw trace file that allows reading individual event records only in a sequential manner, EARL offers random access to individual events and two different categories of abstractions: (i) *state sequences*, and (ii) *pointer attributes*.

State sequences reflect different aspects of the program's overall execution state. The overall execution state consists of a set of (component) states, each of which represents one aspect of the overall state, such as call stacks or message queues. EARL models each component state as a set of events. These sets are stepwise transformed by the sequence of events making up the trace file. That is, an event causes a state transition altering the event set representing the component state by either removing elements and/or adding itself to the set. Thus, for every component state, an event trace defines a *state sequence*. The initial state is always the empty set. Transition rules define how a state is transformed by an event into its successor state. For example, EARL maintains a message queue for every pair of processes. The initial queue is empty. Whenever a send event occurs it is added to the queue, and whenever a receive event occurs the corresponding send event is removed from the queue. Note that the event set representing this component state derives its queue structure from the implicit ordering of events. Other state sequences describe MPI collective communication, OpenMP parallel operations, and lock synchronization, region stacks, and the call tree.

Pointer attributes are event attributes that refer to another related event. For example, receive events provide an attribute called *sendptr* that points to the corresponding send event. The implementation of pointer attributes makes use of state sequences. Other pointer attributes connect matching enter and exit events, link actions on the same lock, and encode call-path information. For a detailed formal definition of the state sequences and pointer attributes provided by EARL, the reader may refer to Wolf [18]. A comprehensive documentation of EARL can be found in Bhatia and Wolf [21].

4.2. Compound events

To analyze an application's performance behavior during a program run, EXPERT walks sequentially through the trace file and tries to match the patterns that have been previously specified in the form of compound events. To illustrate how compound events are specified and detected, Figure 2 shows the time-line view of a situation called *late sender*. Process A waits for a message from process B that is sent a significant time after the receive operation has been started. Therefore, most of the time consumed by the receive operation of process A is actually idle time that could be used more effectively. EXPERT recognizes this pattern by waiting for a receive event to appear in the event stream. After capturing such an event, EXPERT follows pointer attributes computed by EARL (dashed lines in Figure 2) to the enter events of the two communication operations to determine the temporal displacement between these two events (idle time in Figure 2).

Conceptually, a compound event specification consists of three parts: (i) a root declaration, (ii) an instantiation part, (iii) constraints, and (iv) a severity expression. Whenever EXPERT encounters

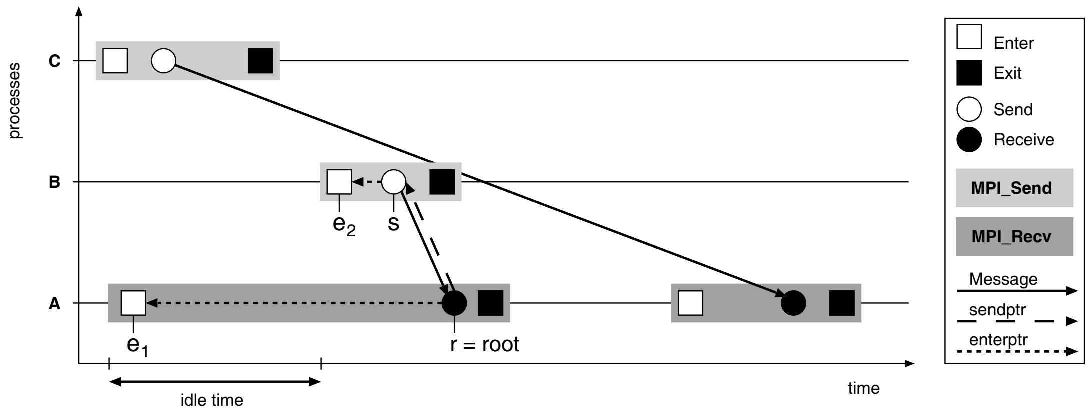


Figure 2. Time-line view of the late-sender compound event.

an instance of the root event type specified in the root declaration, it stops and tries to match the specified compound event. This matching involves two phases. The first phase consists of locating the remaining constituents using state information and pointer attributes, as specified in the instantiation part. The second phase is optional and includes checking additional constraints that need to be satisfied in order to qualify for a match.

A specification of the late-sender situation is given in Figure 3 in a pseudo notation. The instantiation starts at the root event (r), which must be a receive event, and follows pointer attributes to identify the remaining constituents (s, e_1, e_2). The pointer attributes involved are *sendptr*, which points from a receive event to the corresponding send event, and *enterptr*, which points to the enter event on the top of the call stack (i.e. the enter event of the current region instance). The constraints require the communication operations to be of synchronous type with the receive operation being posted earlier than its sending counterpart. Finally, the severity expression calculates the associated execution-time penalty (i.e. the waiting time).

Internally, compound event specifications are written as C++ or Python classes that provide call-back methods to be called upon occurrence of specific event types (root declaration) in the event stream. At present, we maintain two versions of the analyzer: one in C++ for ease of installation and performance and one in Python for design studies. A pattern class registers a call-back method for the root event type, whose instances are later provided as an argument to this method. When being called, the method body performs the instantiation of the compound event along with an optional constraint check. As a result of the trace-access model provided by EARL, the code of the call-back methods can be kept simple by using expressions very similar to the pseudo notation shown in Figure 3. The simplicity is derived from the fact that all higher-level abstractions, such as execution states and links between related events, are expressed in terms of event sets or event references, thus never leaving the familiar notion of an event.

Revisiting the late-sender situation depicted in Figure 2 in the context of the other message sent from process C to A leads to the conclusion that the late-sender pattern could have been avoided or at least

```
ROOT
                  RECV $r$;
INSTANTIATION
                  $s$     $:= r.sendptr$;
                  $e_1$   $:= r.enterptr$;
                  $e_2$   $:= s.enterptr$;
CONSTRAINT
                  $e_1.region$  $=$ MPI_Recv   $\wedge$
                  $e_2.region$  $=$ MPI_Send   $\wedge$
                  $e_1.time$    $< e_2.time$
SEVERITY
                  $e_2.time - e_1.time$
```

Figure 3. Formal specification of late sender.

alleviated by reversing the acceptance order of these two messages. As the message from C is sent earlier than that from B, it will in all probability have reached process A earlier. So instead of waiting for the message from B, A could have used the time better by accepting the message from C first. The late-sender pattern in this context is called *late-sender/wrong-order*. EXPERT recognizes this situation by examining the execution state computed by EARL at the moment when A receives the message from B. It inspects the queue of messages (i.e. their send events) sent to A and checks whether there are older messages than the message that has just been received. In the figure, the queue would contain the event of sending the message from C to A.

As our example suggests, the patterns recognized by EXPERT are organized in a specialization hierarchy, as shown in Figure 4, with patterns referring to rather general performance properties at the top and more specific properties at the bottom. There are two types of patterns: (i) simple profiling patterns (white) aggregating the time spent in certain MPI calls or code regions, and (ii) patterns describing complex inefficiency situations (gray) usually described by more complex compound events (e.g. late sender in point-to-point communication or synchronization delay before all-to-all operations). There are complex patterns defined for MPI-1, MPI-2, SHMEM, and OpenMP. A description of the patterns supported so far can be found in [18,22]. Hardware-counter readings are merely aggregated and can be considered as a special case of profiling patterns augmenting the time-based metrics [23].

With the exception of hardware-counter patterns, each pattern calculates a (call path, location) matrix containing the time incurred by the application as a result of a specific property in a particular (call path, location) pair, where a location is a process or thread. Thus, EXPERT maps the (performance property, call path, location) space onto the losses caused by a particular performance property, while the program was executing in a particular call path at a particular location. After finishing the analysis, the mapping is written to a file and can be viewed using the CUBE display tool.

Figure 5 shows the results for an MPI application called TRACE [24], which simulates the subsurface water flow in variably saturated porous media. The application was executed with 32 processes on a Linux cluster with eight Pentium III Xeon (550 MHz) 4-way nodes. To reduce the size of the trace file, the instrumentation of user functions was restricted to the main solver routine

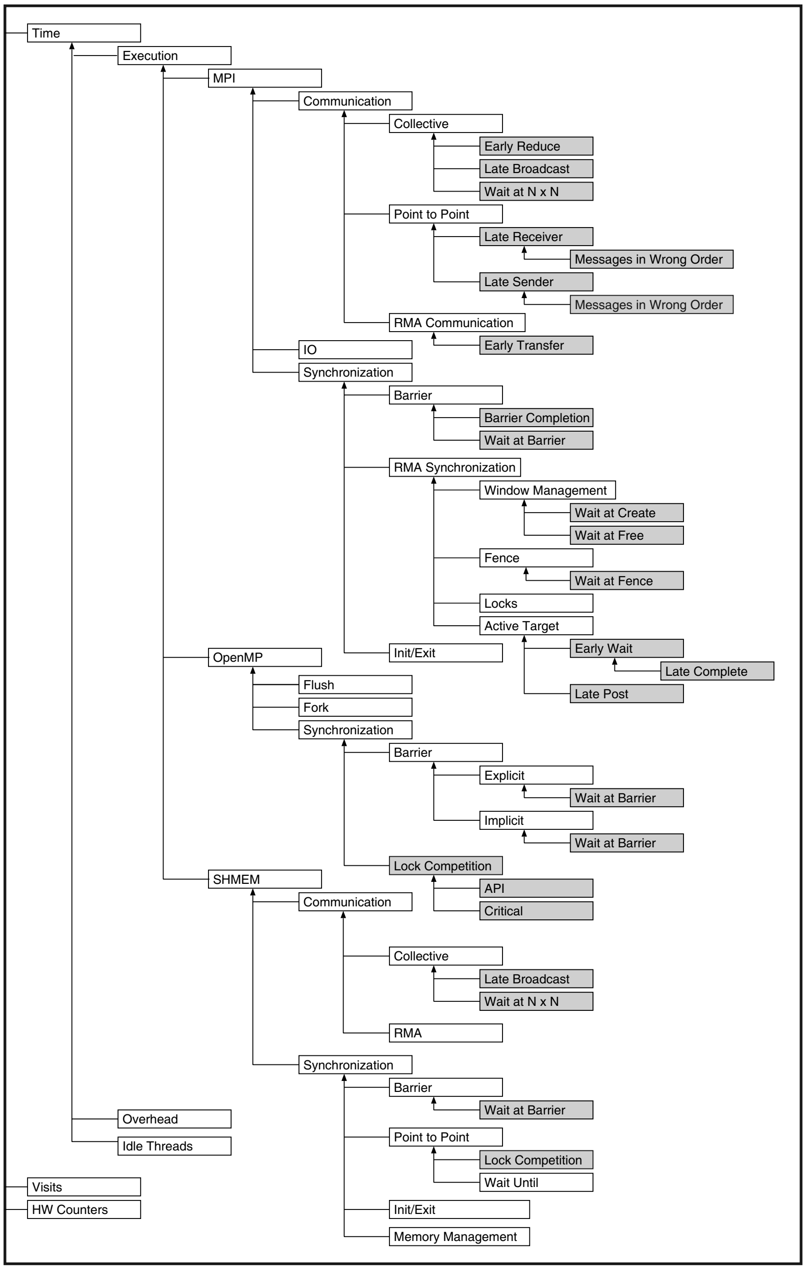
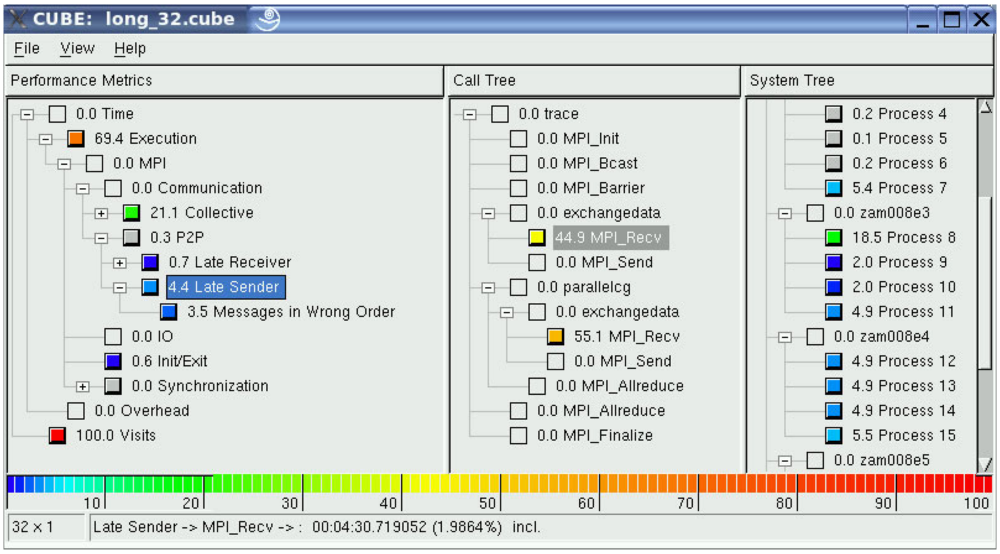


Figure 4. EXPERT pattern hierarchy.



```
CUBE: long_32.cube
File  View  Help

Performance Metrics
0.0 Time
  69.4 Execution
    0.0 MPI
      0.0 Communication
        21.1 Collective
        0.3 P2P
          0.7 Late Receiver
          4.4 Late Sender
            3.5 Messages in Wrong Order
      0.0 IO
      0.6 Init/Exit
      0.0 Synchronization
  0.0 Overhead
100.0 Visits

Call Tree
0.0 trace
  0.0 MPI_Init
  0.0 MPI_Bcast
  0.0 MPI_Barrier
  0.0 exchangedata
    44.9 MPI_Recv
    0.0 MPI_Send
  0.0 parallelcg
    0.0 exchangedata
      55.1 MPI_Recv
      0.0 MPI_Send
    0.0 MPI_Allreduce
  0.0 MPI_Allreduce
  0.0 MPI_Finalize

System Tree
    0.2 Process 4
    0.1 Process 5
    0.2 Process 6
    5.4 Process 7
  0.0 zam008e3
    18.5 Process 8
    2.0 Process 9
    2.0 Process 10
    4.9 Process 11
  0.0 zam008e4
    4.9 Process 12
    4.9 Process 13
    4.9 Process 14
    5.5 Process 15
  0.0 zam008e5

10  20  30  40  50  60  70  80  90  100
32 x 1   Late Sender -> MPI_Recv -> :  00:04:30.719052 (1.9864%)  incl.
```

Figure 5. Performance results for MPI application TRACE.

`parallelcg()` and a function responsible for exchanging buffer contents used inside and outside the main solver (`exchangedata()`).

Almost 8% (4.4% + 3.5%) of the overall execution time was spent waiting in a late-sender situation caused by `exchangedata()`, and a little less than half of it can be attributed to the main solver's operation. However, when executing not on behalf of `parallelcg()`, `exchangedata()` called the receive operation only a few times, pointing to a small number of larger late-sender instances. Also, a significant fraction of the overall late-sender time (3.5% of execution time) could be classified as the more specialized wrong-order pattern.

To ensure efficiency and allow for more compact pattern specifications, the search process takes advantage of the specialization relationships existing between different patterns in a stepwise refinement process [4]. Pattern classes register not only for primitive events, i.e. events as they appear in the event stream, but also register for compound events detected by others and publish compound events that they themselves detect. Transferred to our example, the simple late-sender class publishes all pattern instances it detects. Conversely, the class describing the combined situation late-sender/wrong-order registers for these instances and then, upon receiving such an instance, only needs to check the message queue, as described previously. The benefit is twofold: a more compact specification, as the late-sender part of the pattern specification need not to be repeated, and a reduction of work because the matching of the simple late-sender is performed only once.

EXPERT is designed in a modular fashion, separating the pattern specifications from the actual analysis process. Internally, the semantics of individual pattern classes are hidden behind a common class interface, which makes it easy to modify existing patterns or to extend the current pattern base, for example in order to integrate application-specific patterns, a feature beneficially utilized in Section 5.

5. VIRTUAL TOPOLOGIES

In many parallel applications, each process (or thread) communicates only with a limited number of other processes. For example, a simulation modeling the spread of pollutants in the environment might decompose the overall simulation domain into smaller pieces and assign each of them to a single process. Given this distribution, a process would then only communicate with processes owning subdomains adjacent to its own. The mapping of data onto processes and the neighborhood relationship resulting from this mapping is called a *virtual topology*. In general, a virtual topology is specified as a graph. Many applications use Cartesian topologies, such as two- or three-dimensional grids. Virtual topologies can include processes or threads, depending on the programming model being used. We argue that topological knowledge can help identify performance problems more effectively, especially as many parallel algorithms are parameterized in terms of a virtual topology and this topology often influences the order in which certain computations are performed.

In Section 4, we demonstrated that automatic pattern analysis in event traces can help generate high-level feedback on an application's performance. We identified wait states, i.e. intervals during which a process has to wait (e.g. for a message to arrive), recognizable by temporal displacements between individual events across multiple control flows but without utilizing any information on logical adjacency between processes or threads. We now show that enriching the information contained in event traces with topological knowledge significantly raises the abstraction level of the feedback returned. In particular, we demonstrate that topological information permits the use of the following techniques: (i) identifying higher-level events related to distinct phases of the parallelization scheme applied in an application in order to refine the present classification of wait states targeted by our pattern analysis, and (ii) exposing correlations of these wait states with the topological characteristics of affected processes by visually mapping their severity onto the virtual topology.

As an example, we show correlations between late-sender instances and certain phases in wavefront algorithms, a popular parallelization scheme used to solve particle transport problems. KOJAK has recently been made capable of recording topological information as part of the event trace and of visualizing the severity of the analyzed behaviors in a topological display [25]. To keep this extension simple, we restricted ourselves to Cartesian topologies as a common case found in many of today's parallel applications.

The benchmark code SWEEP3D [26] is an MPI program performing the core computation of a real ASCI application. It calculates the flux of neutrons through each cell of a three-dimensional grid (i, j, k) along several possible directions (angles) of travel. The angles are split into eight octants, each corresponding to one of the eight directed diagonals of the grid. To exploit parallelism, SWEEP3D maps the (i, j) planes of the three-dimensional domain onto a two-dimensional grid of processes. The parallel computation follows a pipelined wavefront process that propagates data along diagonal lines through the grid. Figure 6 shows the data-dependence graph for a 3×3 grid. The long arrows symbolize data dependencies, while diagonal lines cut through algorithmically independent processes and represent the computation as it progresses in the form of 'wavefronts' from the lower left to the upper right corner (short arrows). The actual direction of the wavefront is determined by the

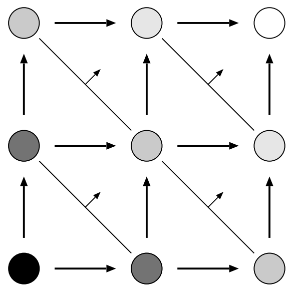


Figure 6. SWEEP3D wavefront propagation.

particular angle or octant being processed at a given moment. The code initiates wavefronts from all four corners of the two-dimensional grid of processes. The wavefronts are pipelined to enable multiple wavefronts to follow each other along the same direction simultaneously. Performance models of wavefront processes, in particular as they appear in SWEEP3D, have been extensively studied [27,28]. With EXPERT we analyze the characteristics of wavefront communication from an experimental viewpoint with an emphasis on wait states resulting from the data dependencies illustrated in Figure 6.

Although parallel operation in SWEEP3D can be very efficient once the pipeline is filled, the opportunity for parallelism is limited whenever the direction of the wavefront changes and the pipeline has to be refilled, although the algorithm allows for some overlap between pipelines in different directions. SWEEP3D uses `MPI_Recv()` calls that are likely to block whenever the pipeline is refilled and the calling process is distant from the pipeline's origin. This phenomenon is a specific instance of the late-sender pattern discussed earlier.

To investigate this type of behavior, we extended the pattern base normally used by our EXPERT analysis tool and added four patterns describing the occurrence of late-sender instances at the moment of a pipeline direction change (i.e. a refill), one pattern for each direction (i.e. South-West, North-West, North-East, and South-East). The direction change is recognized by maintaining a first-in-first out queue for every process that records the directions of messages received. For this purpose, the direction of every message is calculated using topological information. Since the wavefronts propagate along diagonal lines, as depicted in Figure 6, each wavefront direction has a horizontal as well as a vertical component, involving messages in two different orthogonal directions. We therefore need to consider two potential wait states at the moment of a direction change, each resulting from one of the two direction components. However, special attention has to be paid to processes located at the border of the grid (Figure 6). Because they have a smaller number of neighbors, their inbound as well as their outbound communication may be restricted to one direction only, depending on their position relative to the wavefront propagation.

To validate our design, we chose a problem size of $512 \times 512 \times 150$ grid points and ran the application with 64 processes on a Solaris cluster equipped with UltraSPARC-III 750 MHz processors. The topology was recorded with minimum effort by manually inserting two EPILOG API calls into the module responsible for the domain decomposition. If the application had used MPI topology

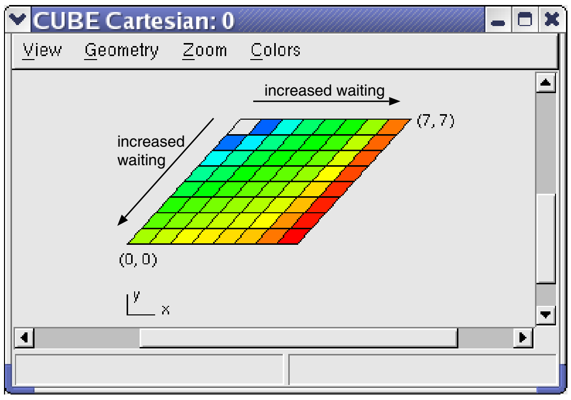


Figure 7. Distribution of late-sender wait states as a result of pipeline refill from the North-West.

support, it would have even been recored completely automatically by appropriate PMPI wrapper functions. The total time spent in late-sender wait states was 25.4%. Late-sender instances observed simultaneously with a pipeline direction change account for about 60% of the overall late-sender time. The times measured for individual directions vary between 6.0% of total execution time for pipeline refill from the North-West and 1.7% for refill from the North-East. Figure 7 shows the CUBE topology view rendering the distribution of late-sender times for pipeline refill from the North-West (i.e. upper left corner). The colors are assigned relative to the maximum and minimum wait times for this particular pattern. As can be seen, the corner reached by the wavefront last incurs most of the waiting times, whereas processes closer to the origin of the wavefront incur less.

This experiment demonstrates that with only minimal user intervention it is possible to automatically highlight a performance problem related to the parallelization scheme applied in SWEEP3D. Manual instrumentation of the direction change is not required. Note that our patterns do not make any assumption about the specifics of the computation performed, and should therefore be applicable to a broad range of wavefront applications. Moreover, although the current implementation applies to wavefront processes based on a two-dimensional domain decomposition, we expect that it can be easily adapted to a three-dimensional decomposition by considering wavefronts propagating along three orthogonal direction components instead of two.

6. CONCLUSION

Because many performance problems of parallel applications involve behavioral dependencies between concurrent control flows, trace analysis is an effective way of identifying undesired wait states that are not obvious without considering temporal and spatial relationships between runtime events across different processes and threads. Automating this process and providing high-level feedback on an application's performance can increase programmer productivity and reduce the time taken to obtain the solution by reducing both development time and execution time.

We have shown that by selecting appropriate abstraction mechanisms, complex performance problems can be specified as patterns in a way that allows the automatic recognition of problem instances in event traces. Using this method, we have found evidence of wait states resulting from an inefficient use of the parallel programming model in real applications. Moreover, such wait states can be correlated with distinct phases of the parallelization strategy applied in a program by utilizing knowledge of the virtual process topology. The modular design of our detection tool allowed us to easily extend the base of predefined patterns and to demonstrate this correlation for wavefront algorithms using algorithm-specific patterns.

While our approach gives automatic performance feedback on a significantly higher level than traditional tools, its dependence on the collection of trace files constrains its scalability on present and future architectures consisting of thousands of processors. Therefore, our future research in this area will focus on using selective instrumentation to record only relevant sections of program execution and on applying parallel and distributed approaches to the processing and reduction of trace data.

REFERENCES

1. Nagel W, Weber M, Hoppe H-C, Solchenbach K. VAMPIR: Visualization and analysis of MPI resources. *Supercomputer* 1996; **12**(1):69–80.
2. Labarta J, Girona S, Pillet V, Cortes T, Gregoris L. DiP: A parallel program development environment. *Proceedings of the 2nd International Euro-Par Conference on Parallel Processing*, Lyon, France, August 1996, vol. 2. Springer: London, 1996.
3. Wolf F, Mohr B. Automatic performance analysis of hybrid MPI/OpenMP applications. *Journal of Systems Architecture (Special Issue on Evolutions in Parallel Distributed and Network-based Processing)* 2003; **49**(10–11):421–439.
4. Wolf F, Mohr B, Dongarra J, Moore S. Efficient pattern search in large traces through successive refinement. *Proceedings of the European Conference on Parallel Computing (Euro-Par)*, Pisa, Italy, August–September 2004 (*Lecture Notes in Computer Science*, vol. 3149). Springer: Berlin, 2004; 47–54.
5. Fahringer T, Gerndt M, Mohr B, Wolf F, Riley G, Träff JL. Knowledge specification for automatic performance analysis. *Technical Report FZJ-ZAM-IB-2001-08*, ESPRIT IV Working Group APART, Forschungszentrum Jülich, August 2001. Revised version.
6. Fahringer T, Seragiotto Júnior C. Modelling and detecting performance problems for distributed and parallel programs with JavaPSL. *Proceedings of the 2001 International Conference on Supercomputers (SC2001)*, Denver, CO, November 2001. ACM/IEEE Computer Society Press: Los Alamitos, CA, 2001.
7. Fürlinger K, Gerndt M. Distributed application monitoring for clustered SMP architectures. *Proceedings of the 9th International Euro-Par Conference*, Klagenfurt, Austria, August 2003. Springer: London, 2003.
8. Jorba J, Margalef T, Luque E. Performance analysis of parallel applications with KappaPI 2. *Proceedings of Parallel Computing 2005 (ParCo 2005)* Malaga, Spain, September 2005 (*NIC Series*, vol. 33). John von Neumann Institute for Computing: Jülich, 2005.
9. Vetter J. Performance analysis of distributed applications using automatic classification of communication inefficiencies. *Proceedings of the 14th International Conference on Supercomputing*, Santa Fe, NM, May 2000; 245–254.
10. Bates PC. Debugging programs in a distributed system environment. *PhD Thesis*, University of Massachusetts, February 1986.
11. Miller BP, Callaghan MD, Cargille JM, Hollingsworth JK, Irvine RB, Karavanic KL, Kunchithapadam K, Newhall T. The Paradyn parallel performance measurement tool. *IEEE Computer* 1995; **28**(11):37–46.
12. Ahn DH, Vetter JS. Scalable analysis techniques for microprocessor performance counter metrics. *Proceedings of the Conference on Supercomputers (SC2002)*, Baltimore, MD, November 2002. ACM/IEEE Computer Society Press: Los Alamitos, CA, 2002.
13. Huband S, McDonald C. A preliminary topological debugger for MPI programs. *Proceedings of the 1st IEEE/ACM International Symposium on Cluster Computing and the Grid*, Buyya R, Mohay G, Roe P (eds.). IEEE Computer Society Press: Los Alamitos, CA, 2001; 422–429.

14. Browne S, Dongarra J, Garner N, Ho G, Mucci P. A portable programming interface for performance evaluation on modern processors. *The International Journal of High Performance Computing Applications* 2000; **14**(3):189–204.
15. Malony AD, Shende S. Performance technology for complex parallel and distributed systems. *Quality of Parallel and Distributed Programs and Systems*, Kacsuk P, Kotsis G (eds.). Nova Science: New York, 2003; 25–41.
16. DeRose L, Hoover T Jr, Hollingsworth JK. The dynamic probe class library—an infrastructure for developing instrumentation for performance tools. *Proceedings of the International Parallel and Distributed Processing Symposium*, April 2001. IEEE Computer Society Press: Los Alamitos, CA, 2001.
17. Mohr B, Malony A, Shende S, Wolf F. Design and prototype of a performance tool interface for OpenMP. *The Journal of Supercomputing* 2002; **23**(August):105–128.
18. Wolf F. Automatic performance analysis on parallel computers with SMP nodes. *PhD Thesis*, RWTH Aachen, Forschungszentrum Jülich, February 2003.
19. Rabenseifner R. The controlled logical clock—a global time for trace based software, monitoring of parallel applications in workstation clusters. *Proceedings of the 5th EUROMICRO Workshop on Parallel and Distributed Processing (PDP)*, London, January 1997. IEEE Computer Society Press: Los Alamitos, CA, 1997; 477–484.
20. Song F, Wolf F, Bhatia N, Dongarra J, Moore S. An algebra for cross-experiment performance analysis. *Proceedings of the International Conference on Parallel Processing (ICPP)*, Montreal, Canada, August 2004. IEEE Computer Society Press: Los Alamitos, CA, 2004.
21. Bhatia N, Wolf F. EARL–API documentation. *Technical Report ICL-UT-04-03*, University of Tennessee, Innovative Computing Laboratory, October 2004.
22. Mohr B, Kühnal A, Hermanns M-A, Wolf F. Performance analysis of one-sided communication mechanisms. *Mini-Symposium on Tools Support for Parallel Programming*. *Proceedings of Parallel Computing 2005 (ParCo 2005)*, Malaga, Spain, September 2005 (*NIC Series*, vol. 33). John von Neumann Institute for Computing: Jülich, 2005.
23. Wylie B, Mohr B, Wolf F. Holistic hardware counter performance analysis of parallel programs. *Proceedings of Parallel Computing 2005 (ParCo 2005)*, Malaga, Spain, September 2005 (*NIC Series*, vol. 33). John von Neumann Institute for Computing: Jülich, 2005.
24. Forschungszeatrum Jülich. Solute transport in heterogeneous soil–aquifer systems. http://www.fz-juelich.de/icg/icg-iv/modeling [7 September 2006].
25. Bhatia N, Song F, Wolf F, Mohr B, Dongarra J, Moore S. Automatic experimental analysis of communication patterns in virtual topologies. *Proceedings of the International Conference on Parallel Processing (ICPP)*, Oslo, Norway, June 2005. IEEE Computer Society Press: Los Alamitos, CA, 2005.
26. Accelerated Strategic Computing Initiative (ASCI). The ASCI sweep3d Benchmark Code. http://www.llnl.gov/asci_benchmarks/ [7 September 2006].
27. Hoisie A, Lubeck O, Wasserman H. Performance analysis of wavefront algorithms on very-large scale distributed systems. *Proceedings of the Workshop on Wide Area Networks and High Performance Computing* (*Lectures Notes in Control and Information Sciences*, vol. 249). Springer: London, 1998; 171–187.
28. Sundaram-Stukel D, Vernon MK. Predictive analysis of a wavefront application using LogGP. *Proceedings of the 7th ACM SIGPLAN Symposium on Principles and Practices of Parallel Programming (PPoPP '99)*, Atlanta, GA, May 1999. *ACM SIGPLAN Notices* 1999; **34**(8):141–150.