Building Your Own Performance Tools

Sameer Shende

Department of Computer and Information Science,

University of Oregon

sameer@cs.uoregon.edu

http://www.cs.uoregon.edu/research/paracomp/tau

Motivation

To discuss issues in instrumentation, measurement, analysis that highlight the choices available for building new tools for evaluating the performance of applications.

Outline

- Introduction to performance evaluation
- Instrumentation techniques
- Measurement techniques
  - Profiling
  - Tracing, synchronization issues
- Analysis techniques
  - Visualization of performance data
- Problems
- Conclusions

Introduction

- Understanding the behavior of parallel programs
  - Performance profiling: What is the relative contribution of routines?
  - Tracing: When do events take place?
  - Bottleneck detection: Where do bottlenecks lie?
  - Debugging: How can I correct the problem?

Understanding Application Performance

- **instrumentation** or modification of the program to generate performance data
- **measurement** of interesting aspects of execution
- **analysis** of the performance data.

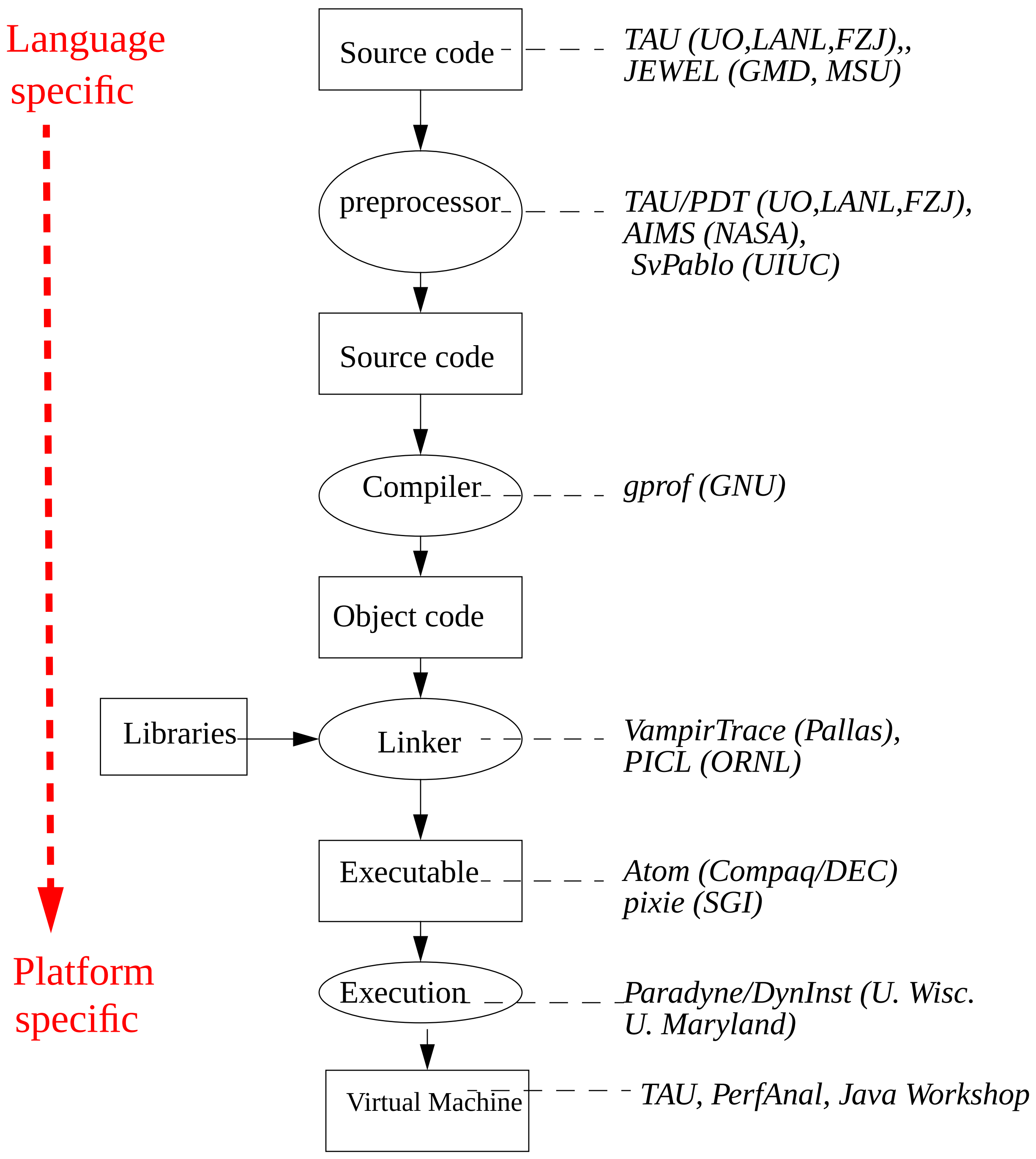
Instrumentation

- **What** is an event?
- **When** does an event get triggered?
- **How** do we add instrumentation to the program?

When does an event get triggered?

- When some point is reached during an execution
  - breakpoint/watchpoint
  - synchronization operation
  - routine entry/exit
- When some internal condition is satisfied
  - Interrupt (time/counts) for sampling
- When some external condition is satisfied
  - signal by debugger/user

When is instrumentation added?


Language specific
Platform specific
Source code
TAU (UO,LANL,FZJ),, JEWEL (GMD, MSU)
preprocessor
TAU/PDT (UO,LANL,FZJ), AIMS (NASA), SvPablo (UIUC)
Source code
Compiler
gprof (GNU)
Object code
Libraries
Linker
VampirTrace (Pallas), PICL (ORNL)
Executable
Atom (Compaq/DEC) pixie (SGI)
Execution
Paradyne/DynInst (U. Wisc. U. Maryland)
Virtual Machine
TAU, PerfAnal, Java Workshop

Instrumentation Approaches

- Manual source code instrumentation
  - Recompile the application
  - Instrumentation API
- Preprocessor
  - Source -to- source transformation
  - Requires a parser for each language
- Compiler
  - Access to code mappings
- Library level instrumentation
  - Interposition libraries: wrappers & callbacks
- Binary Instrumentation
  - Binary rewriting
  - Runtime instrumentation
- Virtual Machine instrumentation

Instrumentation

- ❑ Which is the best instrumentation approach?
- ❑ Is a combination better in some cases?
- ❑ Simplicity vs. Flexibility?

Multi-Level Instrumentation : Example

- Multi-language applications (Java, C++, C, Fortran)
- Hybrid execution models (Java threads, MPI)
- JNI/native Java implementations of MPI Java Interface
  - Java Virtual Machine Profiler Interface (JVMPI)
  - Java Native Interface (JNI)
  - MPI Profiling Interlace

Java Virtual Machine Profiler Interface (JVMPI)

- ❑ Profiling Hooks into the Virtual Machine
- ❑ In-process profiling agent instruments Java application
- ❑ No changes to the Java source code, bytecode, or the executable code of the JVM
- ❑ Two-way call interface
- ❑ Profiler agent is a shared object (libTAU.so) loaded at runtime
- ❑ Agent registers events to the JVMPI
- ❑ JVMPI notifies events to the agent at runtime
- ❑ Agent uses JNI to invoke JVMPI control routines

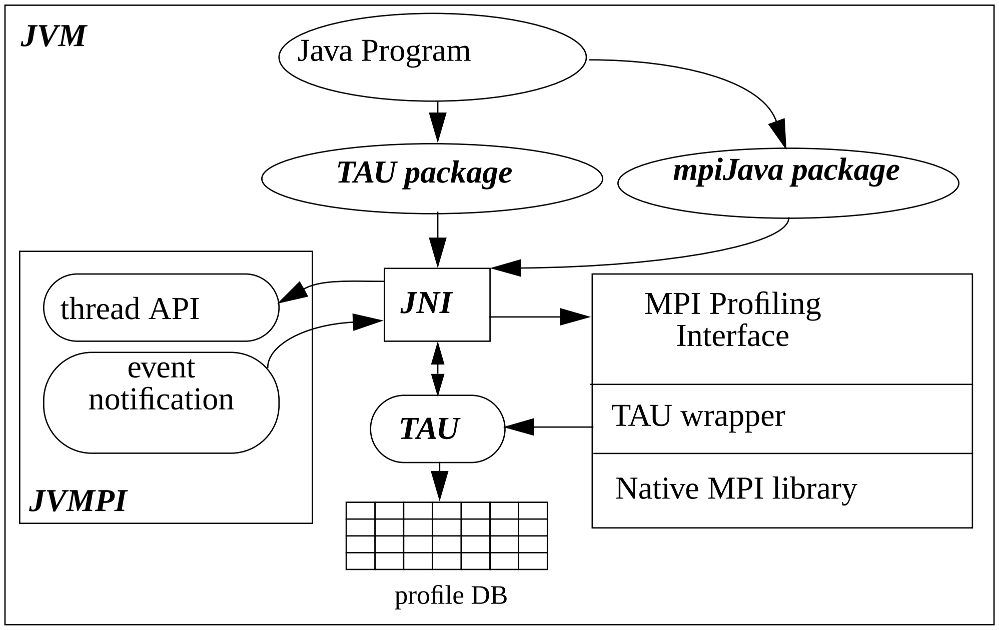
JVMPI Events

- ❑ Method transition events triggered at method entry and exits
- ❑ Memory events triggered when an object is allocated, moved, or deleted
- ❑ Heap arena events triggered when an arena is created or destroyed
- ❑ Garbage collection start and finish events
- ❑ Loading and unloading in memory events for classes and compiled methods
- ❑ JNI global and weak global reference allocation and deallocation events
- ❑ Monitor events for contended Java and raw monitors triggered when a thread attempts to enter, actually enters, or exits a monitor that is accessed by more than one thread
- ❑ Monitor wait events triggered when a thread is about to wait or finishes waiting on an object
- ❑ Thread start and end events when a thread starts or stops executing in the virtual machine
- ❑ Events that request a dump or resetting of the profiling data gathered by the in-process profiling agent
- ❑ Virtual machine initialization and shutdown events

Agent JVMPI interaction

- ❑ create a daemon thread in the virtual machine
- ❑ enable or disable the notification of an event
- ❑ enable, disable or force a garbage collection in the virtual machine
- ❑ obtain information regarding the current method call stack trace for a given thread
- ❑ obtain the accumulated CPU time consumed by the current thread
- ❑ obtain information about the object where a method took place
- ❑ get or set a pointer-sized thread-local storage data structure that can be used to record per-thread profiling data
- ❑ create or destroy a raw monitor. Raw monitors are not associated with Java objects and can be used by the profiler agent to maintain consistency of multi-threaded profiling data
- ❑ enter, exit or wait on a raw monitor for mutual exclusion. It can also notify all threads that are waiting on a raw monitor or specify a time-out period while waiting
- ❑ resume or suspend a thread
- ❑ exit the virtual machine

Integration of Multi-Level Instrumentation APIs



- Common TAU database for multiple sources

Outline

Measurement : Profiling

- shows summary statistics of performance **metrics**
  - CPU time spent in a routine
  - no. of secondary data cache misses for a statement
  - number of profiled routines invoked by a routine...
- presented as sorted lists showing **contribution** of routines
- implemented by **sampling** or **measured** process timing

Profiling Techniques

- sampling (PC/Callstack)
  - time based (hardware interval timer)

    `prof, gprof (GNU)`
  - hardware performance counters based (after *n* instructions, data cache misses...)

    `SpeedShop (SGI), PCL(FZJ), PAPI (UTK)`

➠ Estimates profile, low overhead

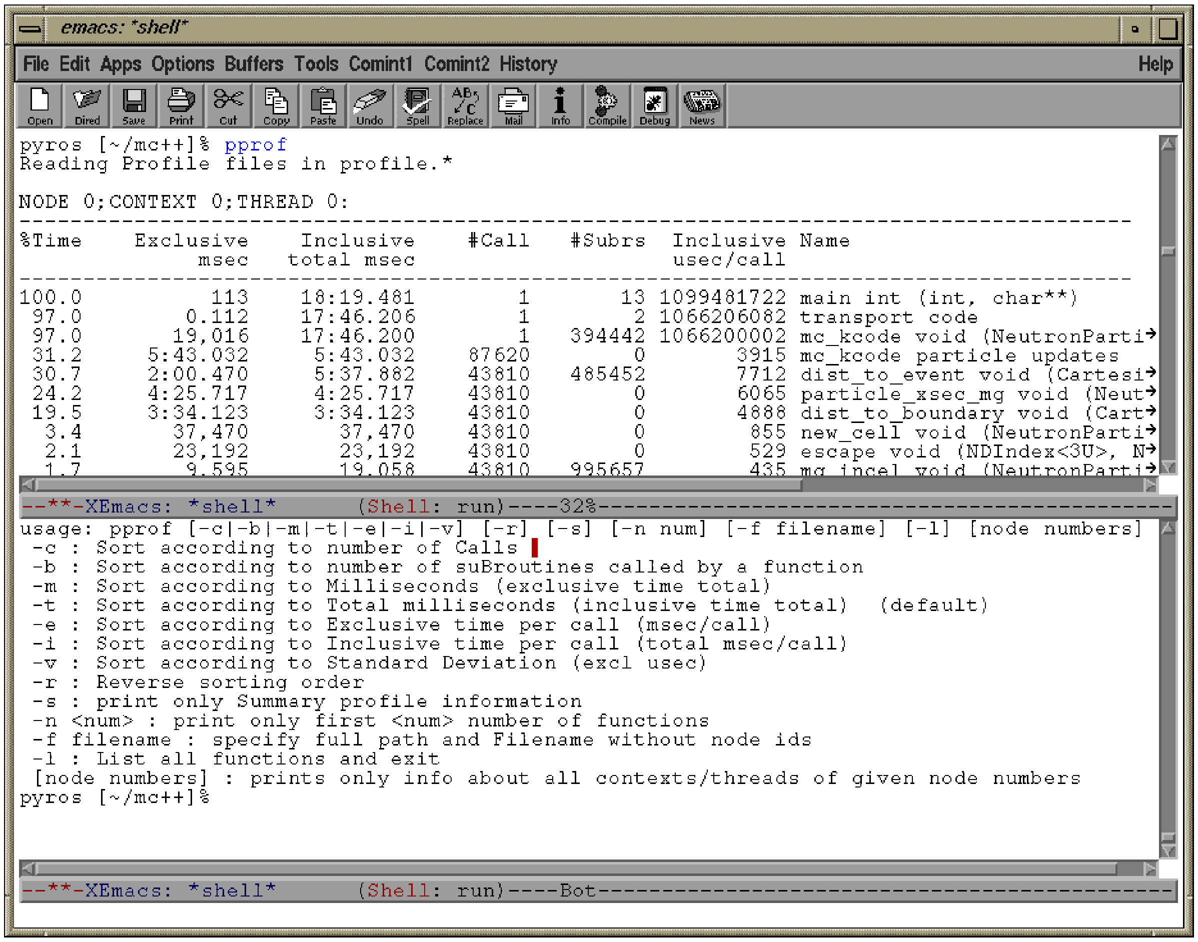
- measured process timing
  - routine entry/exit

    `TAU (U. Oregon, LANL, FZJ)`

➠ Accurate, overhead depends on frequency of invocation

Example of Profiling using TAU

- pprof sorts lists of performance metrics


emacs: *shell*
File Edit Apps Options Buffers Tools Comint1 Comint2 History Help
Open Dired Save Print Cut Copy Paste Undo Spell Replace Mail Info Compile Debug News

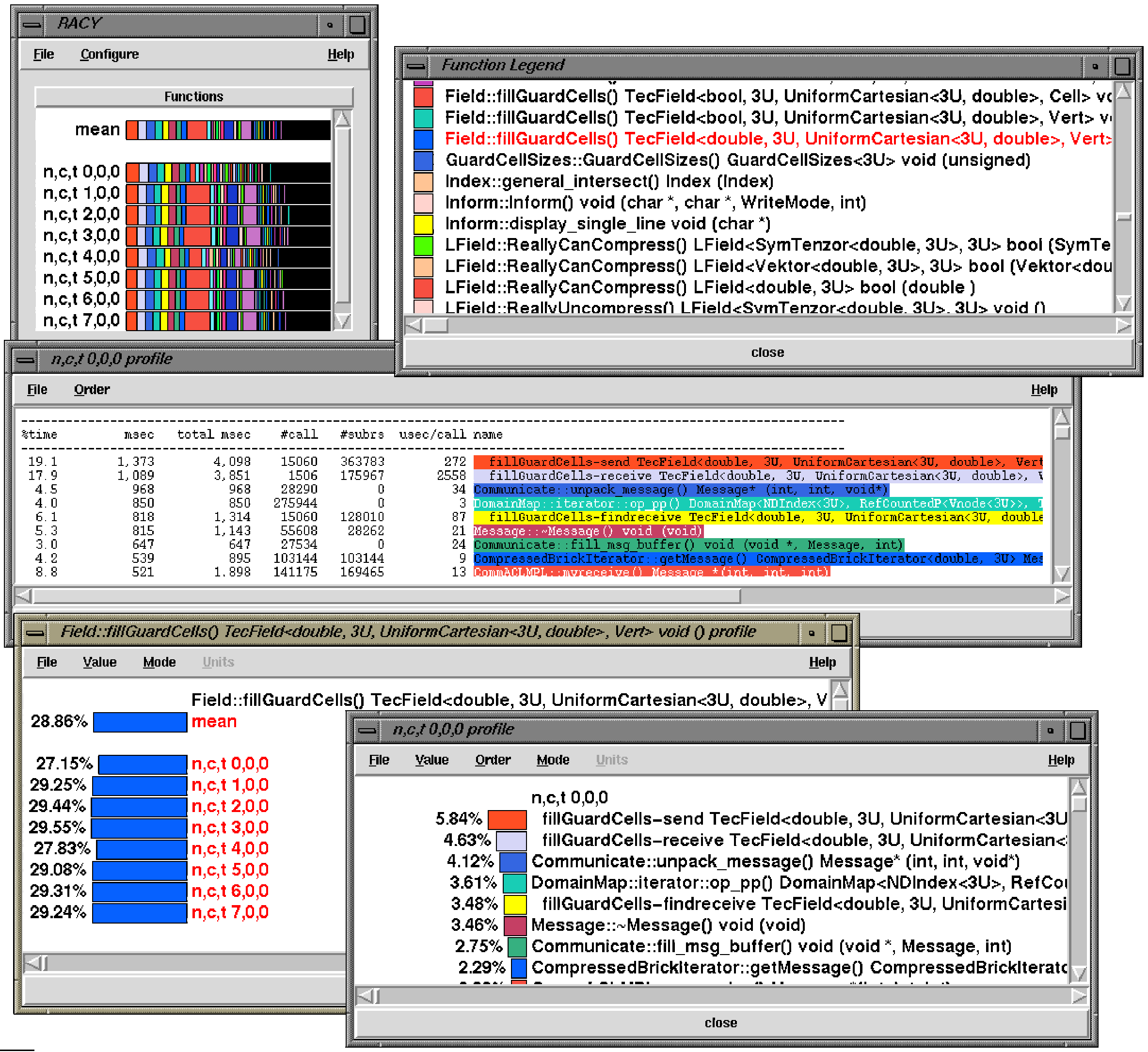
```
pyros [~/mc++]% pprof
Reading Profile files in profile.*

NODE 0;CONTEXT 0;THREAD 0:
---------------------------------------------------------------------------------------
%Time    Exclusive    Inclusive     #Call    #Subrs  Inclusive Name
              msec   total msec                       usec/call
---------------------------------------------------------------------------------------
100.0          113    18:19.481         1        13 1099481722 main int (int, char**)
 97.0        0.112    17:46.206         1         2 1066206082 transport code
 97.0       19,016    17:46.200         1    394442 1066200002 mc_kcode void (NeutronParti→
 31.2     5:43.032     5:43.032     87620         0       3915 mc_kcode particle updates
 30.7     2:00.470     5:37.882     43810    485452       7712 dist_to_event void (Cartesi→
 24.2     4:25.717     4:25.717     43810         0       6065 particle_xsec_mg void (Neut→
 19.5     3:34.123     3:34.123     43810         0       4888 dist_to_boundary void (Cart→
  3.4       37,470       37,470     43810         0        855 new_cell void (NeutronParti→
  2.1       23,192       23,192     43810         0        529 escape void (NDIndex<3U>, N→
  1.7        9.595       19.058     43810    995657        435 mg_incel void (NeutronParti→
--**-XEmacs: *shell*        (Shell: run)----32%------------------------------------------
usage: pprof [-c|-b|-m|-t|-e|-i|-v] [-r] [-s] [-n num] [-f filename] [-l] [node numbers]
 -c : Sort according to number of Calls
 -b : Sort according to number of suBroutines called by a function
 -m : Sort according to Milliseconds (exclusive time total)
 -t : Sort according to Total milliseconds (inclusive time total)  (default)
 -e : Sort according to Exclusive time per call (msec/call)
 -i : Sort according to Inclusive time per call (total msec/call)
 -v : Sort according to Standard Deviation (excl usec)
 -r : Reverse sorting order
 -s : print only Summary profile information
 -n <num> : print only first <num> number of functions
 -f filename : specify full path and Filename without node ids
 -l : List all functions and exit
 [node numbers] : prints only info about all contexts/threads of given node numbers
pyros [~/mc++]%

--**-XEmacs: *shell*        (Shell: run)----Bot------------------------------------------
```

Example: TAU Profiling Package

- RACY



Function Legend:

🟥 Field::fillGuardCells() TecField<bool, 3U, UniformCartesian<3U, double>, Cell> vo
🟦 Field::fillGuardCells() TecField<bool, 3U, UniformCartesian<3U, double>, Vert> vo
🟦 Field::fillGuardCells() TecField<double, 3U, UniformCartesian<3U, double>, Vert>
🟦 GuardCellSizes::GuardCellSizes() GuardCellSizes<3U> void (unsigned)
🟧 Index::general_intersect() Index (Index)
⬜ Inform::Inform() void (char *, char *, WriteMode, int)
🟨 Inform::display_single_line void (char *)
🟩 LField::ReallyCanCompress() LField<SymTenzor<double, 3U>, 3U> bool (SymTe
🟧 LField::ReallyCanCompress() LField<Vektor<double, 3U>, 3U> bool (Vektor<dou
🟥 LField::ReallyCanCompress() LField<double, 3U> bool (double)
⬜ LField::ReallyUncompress() LField<SymTenzor<double, 3U>, 3U> void ()

n,c,t 0,0,0 profile:

```
---------------------------------------------------------------------------------------------------------------
%time        msec   total msec     #call    #subrs  usec/call name
---------------------------------------------------------------------------------------------------------------
 19.1       1,373       4,098     15060    363783        272   fillGuardCells-send TecField<double, 3U, UniformCartesian<3U, double>, Vert
 17.9       1,089       3,851      1506    175967       2558   fillGuardCells-receive TecField<double, 3U, UniformCartesian<3U, double>, V
  4.5         968         968     28290         0         34 Communicate::unpack_message() Message* (int, int, void*)
  4.0         850         850    275944         0          3 DomainMap::iterator::op_pp() DomainMap<NDIndex<3U>, RefCountedP<Vnode<3U>>, T
  6.1         818       1,314     15060    128010         87   fillGuardCells-findreceive TecField<double, 3U, UniformCartesian<3U, double
  5.3         815       1,143     55608     28262         21 Message::~Message() void (void)
  3.0         647         647     27534         0         24 Communicate::fill_msg_buffer() void (void *, Message, int)
  4.2         539         895    103144    103144          9 CompressedBrickIterator::getMessage() CompressedBrickIterator<double, 3U> Mes
  8.8         521       1.898    141175    169465         13 CommACLMPL::myreceive() Message *(int, int, int)
```

Field::fillGuardCells() TecField<double, 3U, UniformCartesian<3U, double>, V

28.86%	mean
27.15%	n,c,t 0,0,0
29.25%	n,c,t 1,0,0
29.44%	n,c,t 2,0,0
29.55%	n,c,t 3,0,0
27.83%	n,c,t 4,0,0
29.08%	n,c,t 5,0,0
29.31%	n,c,t 6,0,0
29.24%	n,c,t 7,0,0

n,c,t 0,0,0

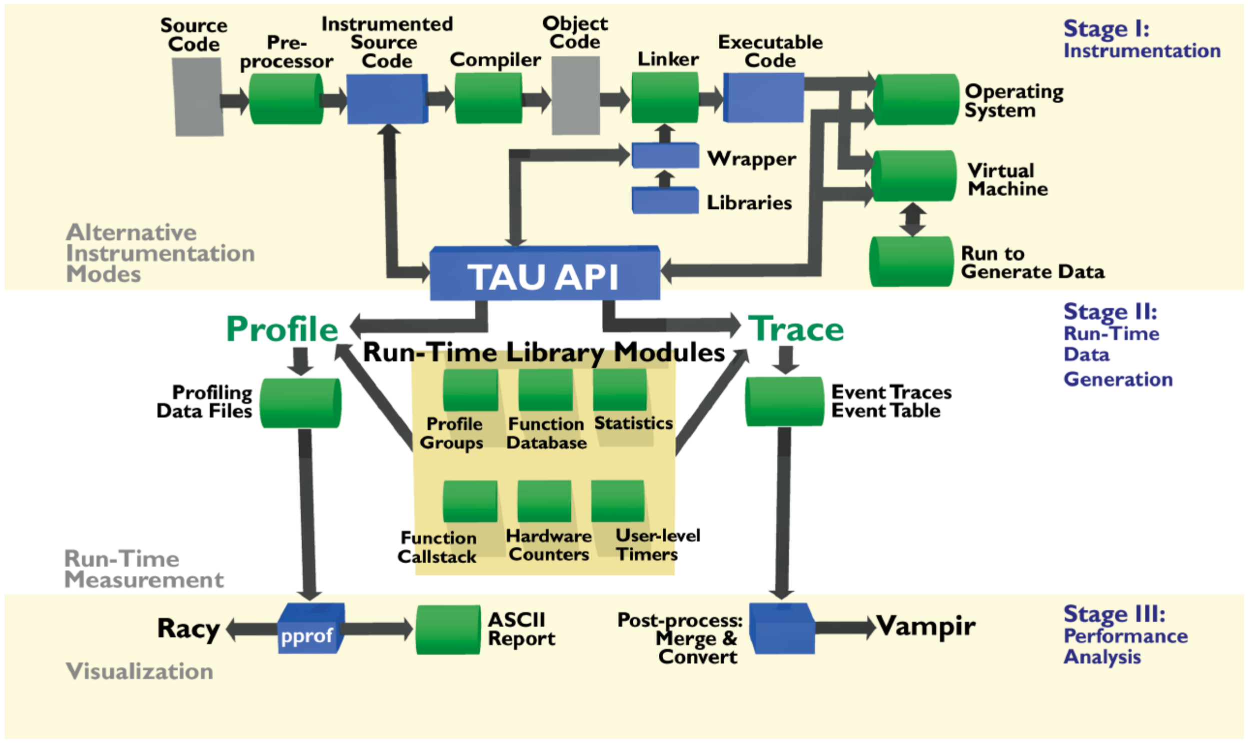
5.84% 🟥	fillGuardCells–send TecField<double, 3U, UniformCartesian<3U
4.63% ⬜	fillGuardCells–receive TecField<double, 3U, UniformCartesian<
4.12% 🟦	Communicate::unpack_message() Message* (int, int, void*)
3.61% 🟦	DomainMap::iterator::op_pp() DomainMap<NDIndex<3U>, RefCou
3.48% 🟨	fillGuardCells–findreceive TecField<double, 3U, UniformCartesi
3.46% 🟥	Message::~Message() void (void)
2.75% 🟩	Communicate::fill_msg_buffer() void (void *, Message, int)
2.29% 🟦	CompressedBrickIterator::getMessage() CompressedBrickIterato

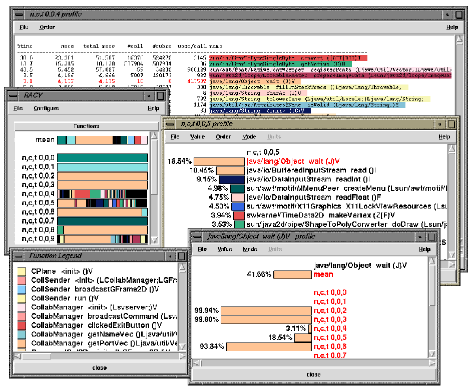
Tracing

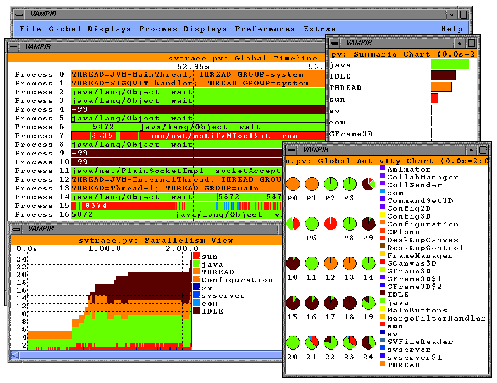
- tracing highlights the temporal aspect of performance variations, showing **when** and **where** in the code performance is achieved
- logging events (routine transitions/messages/user-def.)
  - event identifier
  - timestamp when the event occurred
  - where it occurred (node, context, thread ids)
  - optional field of event specific information
  - plus, event headers (event characteristics)

Architecture of TAU

Tuning and Analysis Utilities







Example

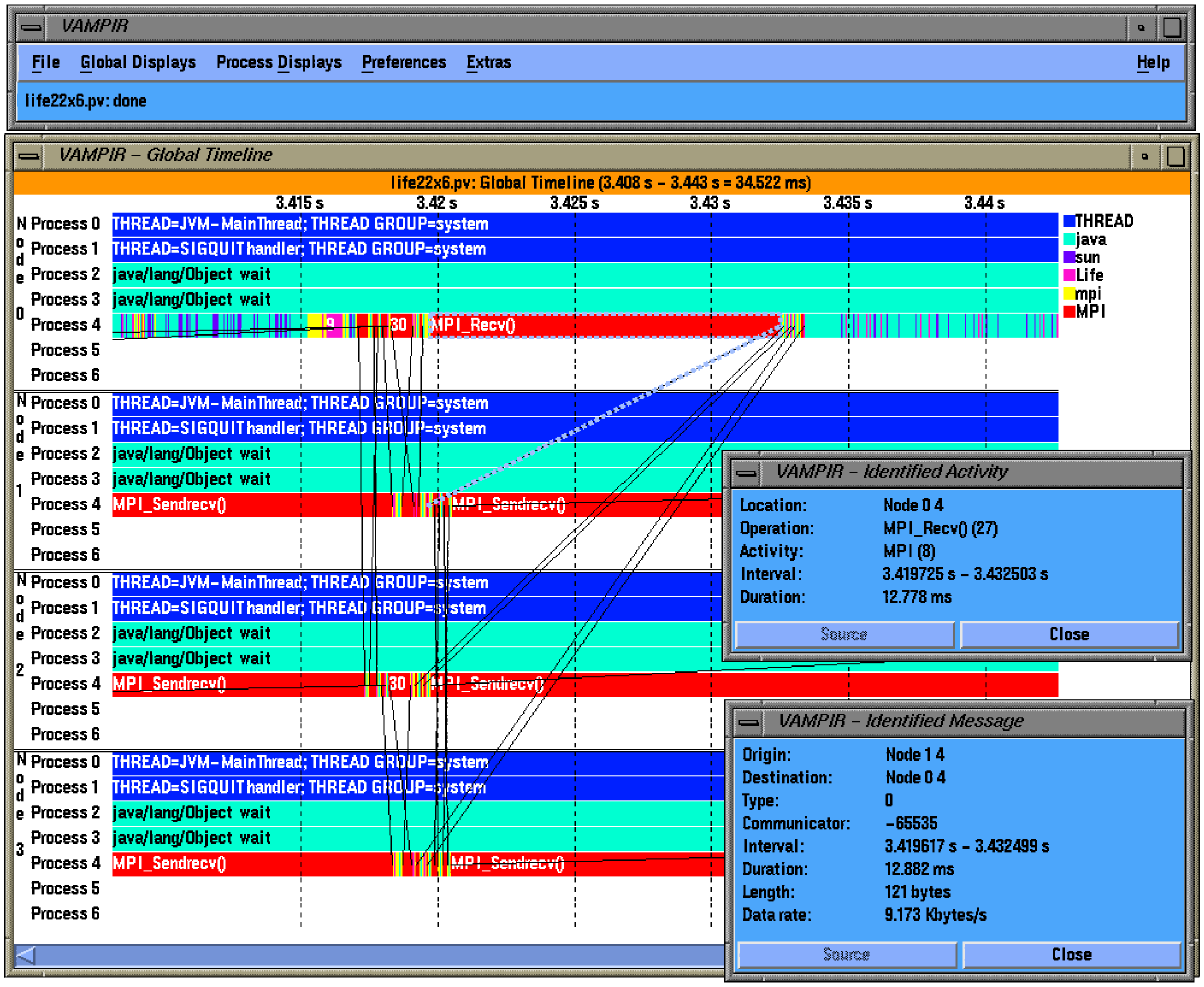
- ❑ Tracing: Visualization in Vampir [http://www.pallas.de]

```
% prunjava 4 Life

% tau_merge tautrace*.trc Life.trc

% tau_convert -vampir Life.trc tau.edf
Life.pv

% vampir Life.pv
```



VAMPIR – Identified Activity

Location:	Node 0 4
Operation:	MPI_Recv() (27)
Activity:	MPI (8)
Interval:	3.419725 s – 3.432503 s
Duration:	12.778 ms

VAMPIR – Identified Message

Origin:	Node 1 4
Destination:	Node 0 4
Type:	0
Communicator:	-65535
Interval:	3.419617 s – 3.432499 s
Duration:	12.882 ms
Length:	121 bytes
Data rate:	9.173 Kbytes/s

Analysis of Performance Data

- ❑ Pablo (U. Illinois, Urbana)
  - ❍ User-directed analysis using performance data transformation modules that are interconnected
  - ❍ http://www-pablo.cs.uiuc.edu/
- ❑ Vampir (FZJ, Pallas GmbH)
  - ❍ Commercial trace visualization tool
  - ❍ http://www.pallas.de
- ❑ ParaGraph (NCSA, UIUC)
  - ❍ Rich set of visualizations, extensible
  - ❍ http://www.ncsa.uiuc.edu/

Problems...

- how do we profile/trace in the presence of
  - optimizations (PETE/C++, ZPL)
  - code transformations (Opus/HPF, Fortran-D)?
- how can we compensate for the perturbation caused by the instrumentation?
- how can we map performance data between layers?
- how can we produce meaningful visualizations that can scale to thousands of processors?
- how can we show performance data at a level of abstraction that the user understands?

Conclusions

- Effective choices
  - instrumentation
  - measurement
  - analysis
- Bridging the “semantic-gap”
- Problems and constraints

Unless tools can present performance data in ways that are meaningful to the user, and are consistent with the user's mental model of abstractions, their success will be limited.