

# The Erlang Runtime System

Erik Stenman



# Table of Contents

Preface .....	1
About this book .....	1
How to read this book .....	1
Erlang .....	1
Acknowledgments .....	2
I: Understanding ERTS .....	5
1. Introducing the Erlang Runtime System .....	7
1.1. ERTS and the Erlang Runtime System .....	7
1.2. How to read this book .....	7
1.3. ERTS .....	7
1.3.1. The Erlang Node (ERTS) .....	7
1.3.2. Layers in the Execution Environment .....	8
1.3.3. Distribution .....	10
1.3.4. The Erlang Compiler .....	10
1.3.5. The Erlang Virtual Machine: BEAM .....	11
1.3.6. Processes .....	11
1.3.7. Scheduling .....	11
1.3.8. The Erlang Tag Scheme .....	11
1.3.9. Memory Handling .....	12
1.3.10. The Interpreter and the Command Line Interface .....	12
1.4. Other Erlang Implementations .....	12
1.4.1. Erlang on Xen .....	12
1.4.2. Erjang .....	13
2. The Compiler .....	15
2.1. Compiling Erlang .....	15
2.2. Generating Intermediate Output .....	17
2.3. Compiler Passes .....	20
2.3.1. Compiler Pass: The Erlang Preprocessor (epp) .....	20
2.3.2. Compiler Pass: Parse Transformations .....	20
2.3.3. Compiler Pass: Linter .....	24
2.3.4. Compiler Pass: Save AST .....	24
2.3.5. Compiler Pass: Expand .....	25
2.3.6. Compiler Pass: Core Erlang .....	25
2.3.7. Compiler Pass: Kernel Erlang .....	25
2.3.8. Compiler Pass: BEAM Code .....	25
2.3.9. Compiler Pass: Native Code .....	25
2.4. Other Compiler Tools .....	25
2.4.1. Leex .....	26
2.4.2. Yecc .....	28
2.5. Syntax Tools and Merl .....	30
2.6. Compiling Elixir .....	30
3. Processes .....	31
3.1. What is a Process? .....	31
3.1.1. Listing Processes from the Shell .....	31
3.1.2. Programmatic Process Probing .....	33

3.1.3. Using the Observer to Inspect Processes .....	36
3.2. Processes Are Just Memory .....	39
3.3. The PCB .....	44
3.4. The Garbage Collector (GC) .....	45
3.5. Mailboxes and Message Passing .....	46
3.5.1. Sending Messages in Parallel .....	46
3.6. Lock Free Message Passing .....	46
3.6.1. Memory Areas for Messages .....	47
3.6.2. Inspecting Message Handling .....	47
3.6.3. The Process of Sending a Message to a Process .....	49
3.6.4. Receiving a Message .....	53
3.6.5. Tuning Message Passing .....	53
3.7. The Process Dictionary .....	54
3.8. Dig In .....	54
4. The Erlang Type System and Tags .....	57
4.1. The Erlang Type System .....	57
4.2. The Tagging Scheme .....	58
4.2.1. Tags for Immediates .....	58
4.2.2. Tags for Boxed Terms .....	59
5. The Erlang Virtual Machine: BEAM .....	63
5.1. Working Memory: A stack machine, it is not .....	63
5.2. Dispatch: Directly Threaded Code .....	66
5.3. Scheduling: Non-preemptive, Reduction counting .....	72
5.4. Memory Management: Garbage Collecting .....	73
5.5. BEAM: it is virtually unreal .....	73
6. Modules and The BEAM File Format (16p) .....	75
6.1. Modules .....	75
6.2. The BEAM File Format .....	75
6.2.1. Atom table chunk .....	77
6.2.2. Export table chunk .....	78
6.2.3. Import table chunk .....	79
6.2.4. Code Chunk .....	79
6.2.5. String table chunk .....	81
6.2.6. Attributes Chunk .....	81
6.2.7. Compilation Information Chunk .....	82
6.2.8. Local Function Table Chunk .....	82
6.2.9. Literal Table Chunk .....	83
6.2.10. Abstract Code Chunk .....	84
6.2.11. Compression and Encryption .....	84
6.2.12. Function Trace Chunk (Obsolete) .....	84
6.2.13. Bringing it all Together .....	84
6.2.14. Compact Term Encoding .....	84
7. Generic BEAM Instructions (25p) .....	87
7.1. Instruction definitions .....	87
7.2. BEAM code listings .....	88
7.3. Calls .....	90
7.4. Stack (and Heap) Management .....	90
7.5. Message Passing .....	91

7.5.1. A Minimal Receive Loop .....	92
7.5.2. A Selective Receive Loop .....	92
7.5.3. A Receive Loop With a Timeout .....	92
7.5.4. The Synchronous Call Trick (aka The Ref Trick) .....	93
8. Different Types of Calls, Linking and Hot Code Loading (5p) .....	95
8.1. Hot Code Loading .....	95
8.2. Code Loading .....	96
9. The BEAM Loader .....	97
9.1. Transforming from Generic to Specific instructions .....	97
9.2. Understanding ops.tab .....	97
9.2.1. Transformations .....	97
9.2.2. Specific instruction .....	99
9.3. Optimizations .....	101
9.3.1. select_val optimizations .....	102
9.3.2. pre-hashing of literals .....	102
10. BEAM Internal Instructions .....	105
11. Scheduling .....	107
11.1. Concurrency, Parallelism, and Preemptive Multitasking .....	107
11.2. Preemptive Multitasking in ERTS Cooperating in C .....	108
11.3. Reductions .....	108
11.3.1. How Many Reductions Will You Get? .....	109
11.3.2. What is a Reduction Really? .....	109
11.4. The Process State (or <i>status</i> ) .....	110
11.5. Process Queues .....	111
11.5.1. The Ready Queue .....	112
11.5.2. Waiting, Timeouts and the Timing Wheel .....	112
11.6. Ports .....	113
11.7. Reductions .....	114
11.8. The Scheduler Loop .....	114
11.9. Load Balancing .....	114
11.9.1. Task Stealing .....	115
11.9.2. Migration .....	115
12. The Memory Subsystem: Stacks, Heaps and Garbage Collection .....	119
12.1. The memory subsystem .....	121
12.2. Different type of memory allocators .....	121
12.2.1. The basic allocator: sys_alloc .....	122
12.2.2. The memory segment allocator: mseg_alloc .....	123
12.2.3. The memory allocator framework: alloc_util .....	123
12.2.4. The temporary allocator: temp_alloc .....	125
12.2.5. The heap allocator: eheap_alloc .....	125
12.2.6. The binary allocator: binary_alloc .....	126
12.2.7. The ETS allocator: ets_alloc .....	126
12.2.8. The driver allocator: driver_alloc .....	126
12.2.9. The short lived allocator: sl_alloc .....	126
12.2.10. The long lived allocator: ll_alloc .....	126
12.2.11. The fixed size allocator: fix_alloc .....	126
12.2.12. The standard allocator: std_alloc .....	126
12.3. TODO: system flags for memory .....	126

12.4. Process Memory .....	126
12.4.1. Term sharing .....	127
12.4.2. Message passing .....	128
12.4.3. Binaries .....	130
12.4.4. Garbage Collection .....	132
12.5. Other interesting memory areas .....	141
12.5.1. The atom table .....	141
13. Advanced data structures (ETS, DETS, Mnesia) .....	143
14. IO, Ports and Networking (10p) .....	145
14.1. Standard IO .....	145
14.2. Ports .....	145
14.2.1. Different types of Ports .....	146
14.3. Distributed Erlang .....	147
14.4. Sockets, UDP and TCP .....	147
15. Distribution .....	149
16. Interfacing C — BIFs NIFs and Linked in Drivers .....	151
17. Native Code .....	153
II: Running ERTS .....	155
18. Tracing .....	157
19. Debugging .....	159
19.1. Preliminary Outline .....	159
19.2. Introduction .....	159
19.3. debugger .....	159
19.4. dbg .....	159
19.5. Redbug .....	159
19.5.1. Installing Redbug .....	159
19.5.2. Using Redbug .....	160
19.6. Crash Dumps .....	164
20. Operation and Maintenance .....	165
20.1. Connecting to the System .....	165
20.2. The Shell .....	166
20.2.1. Configuring Your Shell .....	167
20.2.2. Connecting a Shell to a Node .....	167
20.2.3. Breaking (out or in) .....	172
21. Tweaking the Runtime System .....	173
Appendix A: Building the Erlang Runtime System .....	175
A.1. First Time Build .....	175
A.1.1. Prerequisites .....	175
A.2. Getting the source .....	176
A.3. Building with Kerl .....	176
Appendix B: BEAM Instructions .....	177
B.1. Functions and Labels .....	177
B.1.1. label Lbl .....	177
B.1.2. func_info Module Function Arity .....	177
B.2. Test instructions .....	177
B.2.1. Type tests .....	177
B.2.2. Comparisons .....	178
B.3. Function Calls .....	178

B.3.1. call Arity Label .....	178
B.3.2. call_only Arity Label .....	178
B.3.3. call_last Arity Label Deallocate .....	178
B.3.4. call_ext Arity Destination .....	178
B.3.5. call_ext_only Arity Destination .....	179
B.3.6. call_ext_last Arity Destination Deallocate .....	179
B.3.7. bif0 Bif Reg, bif[1,2] Lbl Bif [Arg,...] Reg .....	179
B.3.8. gc_bif[1-3] Lbl Live Bif [Arg, ...] Reg .....	179
B.3.9. call_fun Arity .....	179
B.3.10. apply Arity .....	179
B.3.11. apply_last Arity Dealloc .....	179
<b>B.4. Stack (and Heap) Management .....</b>	<b>179</b>
B.4.1. allocate StackNeed Live .....	180
B.4.2. allocate_heap StackNeed HeapNeed Live .....	181
B.4.3. allocate_zero StackNeed Live .....	181
B.4.4. allocate_heap_zero StackNeed HeapNeed Live .....	181
B.4.5. test_heap HeapNeed Live .....	181
B.4.6. init N .....	181
B.4.7. deallocate N .....	182
B.4.8. return .....	182
B.4.9. trim N Remaining .....	182
<b>B.5. Moving, extracting, modifying data .....</b>	<b>182</b>
B.5.1. move Source Destination .....	182
B.5.2. get_list Source Head Tail .....	182
B.5.3. get_tuple_element Source Element Destination .....	182
B.5.4. set_tuple_element NewElement Tuple Position .....	182
<b>B.6. Building terms .....</b>	<b>182</b>
B.6.1. put_list Head Tail Destination .....	183
B.6.2. put_tuple Size Destination .....	183
B.6.3. put Value .....	183
B.6.4. make_fun2 LambdaIndex .....	183
<b>B.7. Binary Syntax .....</b>	<b>183</b>
B.7.1. bs_put_integer/5 .....	183
B.7.2. bs_put_binary/5 .....	183
B.7.3. bs_put_float/5 .....	183
B.7.4. bs_put_string/2 .....	183
B.7.5. bs_init2/6 .....	183
B.7.6. bs_add/5 .....	183
B.7.7. bs_start_match2/5 .....	183
B.7.8. bs_get_integer2/7 .....	184
B.7.9. bs_get_float2/7 .....	184
B.7.10. bs_get_binary2/7 .....	184
B.7.11. bs_skip_bits2/5 .....	184
B.7.12. bs_test_tail2/3 .....	184
B.7.13. bs_save2/2 .....	184
B.7.14. bs_restore2/2 .....	184
B.7.15. bs_context_to_binary/1 .....	184
B.7.16. bs_test_unit/3 .....	184

B.7.17. bs_match_string/4 .....	184
B.7.18. bs_init_writable/0 .....	184
B.7.19. bs_append/8 .....	184
B.7.20. bs_private_append/6 .....	184
B.7.21. bs_init_bits/6 .....	185
B.7.22. bs_get_utf8/5 .....	185
B.7.23. bs_skip_utf8/4 .....	185
B.7.24. bs_get_utf16/5 .....	185
B.7.25. bs_skip_utf16/4 .....	185
B.7.26. bs_get_utf32/5 .....	185
B.7.27. bs_skip_utf32/4 .....	185
B.7.28. bs_utf8_size/3 .....	185
B.7.29. bs_put_utf8/3 .....	185
B.7.30. bs_utf16_size/3 .....	185
B.7.31. bs_put_utf16/3 .....	185
B.7.32. bs_put_utf32/3 .....	185
<b>B.8. Floating Point Arithmetic .....</b>	<b>185</b>
B.8.1. fclearerror/0 .....	185
B.8.2. fcheckerror/1 .....	186
B.8.3. fmove/2 .....	186
B.8.4. fconv/2 .....	186
B.8.5. fadd/4 .....	186
B.8.6. fsub/4 .....	186
B.8.7. fmul/4 .....	186
B.8.8. fdiv/4 .....	186
B.8.9. frotate/3 .....	186
<b>B.9. Pattern Matching .....</b>	<b>186</b>
B.9.1. select_val .....	186
B.9.2. select_arity_val .....	186
B.9.3. jump .....	186
<b>B.10. Exception handling .....</b>	<b>186</b>
B.10.1. catch/2 .....	186
B.10.2. catch_end/1 .....	187
B.10.3. badmatch/1 .....	187
B.10.4. if_end/0 .....	187
B.10.5. case_end/1 .....	187
<b>B.11. Meta instructions .....</b>	<b>187</b>
B.11.1. on_load .....	187
B.11.2. line .....	187
<b>B.12. Generic Instructions .....</b>	<b>187</b>
<b>B.13. Specific Instructions .....</b>	<b>198</b>
B.13.1. List of all BEAM Instructions .....	198
<b>Appendix C: Full Code Listings .....</b>	<b>207</b>

# Preface

This book is not about how to write correct and beautiful code, I am assuming that you already know how to do that. This book isn't really about profiling and performance tuning either. Although, there is a chapter in this book on tracing and profiling which can help you find bottlenecks and unnecessary usage of resources. There also is a chapter on performance tuning.

These two chapters are the last chapters in the book, and the whole book is building up to those chapters, but the real goal with this book is to give you all the information, all the gory details, that you need in order to really understand the performance of your Erlang application.

## About this book

For anyone who: Want to tune an Erlang installation. Want to know how to debug VM crashes. Want to improve performance of Erlang applications. Want to understand how Erlang really works. Want to learn how to build your own runtime environment.

If you want to debug the VM If you want to extend the VM If you want to do performance tweaking—jump to the last chapter ... but to really understand that chapter you need to read the book.

## How to read this book

The Erlang RunTime System (ERTS) is a complex system with many interdependent components. It is written in a very portable way so that it can run on anything from a gum-stick computer to the largest multicore system with terabytes of memory. In order to be able to optimize the performance of such a system for your application, you need to not only know your application, but you also need to have a thorough understanding of ERTS itself.

With this knowledge of how ERTS works you will be able to understand how your application behaves when running on ERTS, and you will also be able to find and fix problems with the performance of your application. In the second part of this book we will go through how you successfully run, monitor and scale your ERTS application.

You don't need to be an Erlang programmer to read this book, but you will need some basic understanding of what Erlang is. This following section will give you some Erlang background.

## Erlang

In this section we will look at some basic Erlang concepts that are vital to understanding the rest of the book.

Erlang has been called, especially by one of Erlang's creators Joe Armstrong, a concurrency oriented language. Concurrency is definitely at the heart of Erlang, and to be able to understand how an Erlang system works you need to understand the concurrency model of Erlang.

First of all we need to make a distinction between *concurrency* and *parallelism*. In this book *concurrency* is the concept of having two or more processes that **can** execute independently of each other, this can be done by first executing one process then the other or by interleaving the execution, or by executing the processes in parallel. With *parallel* executions we mean that the processes actually execute at the exact same time by using several physical execution units. Parallelism can be achieved on different levels. Through multiple execution units in the execution pipeline in one core, in several cores on one CPU, by several CPUs in one machine or through several machines.

Erlang uses processes to achieve concurrency. Conceptually Erlang processes are similar to most OS processes, they execute in parallel and can communicate through signals. In practice there is a huge difference in that Erlang processes are much more lightweight than most OS processes. Many other concurrent programming languages call their equivalent to Erlang processes *agents*.

Erlang achieves concurrency by interleaving the execution of processes on the Erlang virtual machine, the BEAM. On a multi-core processor the BEAM can also achieve parallelism by running one scheduler per core and executing one Erlang process per scheduler. The designer of an Erlang system can achieve further parallelism by distributing the system on several computers.

A typical Erlang system (a server or service built in Erlang) consists of a number of Erlang *applications*, corresponding to a directory on disk. Each application is made up of several Erlang *modules* corresponding to files in the directory. Each module contains a number of *functions* and each function is made up of *expressions*.

Since Erlang is a functional language it has no statements, only expressions. Erlang expressions can be combined into an Erlang function. A function takes a number of arguments and returns a value. In [Erlang Code Examples](#) we can see some examples of Erlang expressions and functions.

### *Erlang Code Examples*

```
%% Some Erlang expressions:
```

```
true.  
1+1.  
if (X > Y) -> X; true -> Y end.
```

```
%% An Erlang function:
```

```
max(X, Y) ->  
  if (X > Y) -> X;  
  true    -> Y  
end.
```

Erlang has a number of *built in functions* (or *BIFs*) which are implemented by the VM. This is either for efficiency reasons, like the implementation of lists:append (which could be implemented in Erlang). It could also be to provide some low level functionality which would be hard or impossible to implement in Erlang itself, like list\_to\_atom.

Since Erlang/OTP R13B03 you can also provide your own functions implemented in C by using the *Native Implemented Functions (NIF)* interface.

## Acknowledgments

A big thank you to everyone who has contributed in some way by checking in something:

Roberto Alois  
Dmytro Lytovchenko  
Yoshihiro Tanaka  
Anthony Molinaro  
Alexandre Rodrigues  
Yoshihiro TANAKA  
Ken Causey  
Lukas Larsson  
Kim Shrier  
Ramkumar Rajagopalan  
Greg Baraghimian  
Michał Piotrowski  
Marc van Woerkom  
Anton N Ryabkov  
Yves Müller  
Andrea Leopardi  
techgaun  
Trevor Brown  
Alex Fu  
tomdos  
Alex Jiao  
Amir Moulavi  
Antonio Nikishaev  
Benjamin Tan Wei Hao  
Cameron Price  
Chris Yunker  
Davide Bettio  
Eric Yu  
Erick Dennis  
Jan Lehnardt  
Juan Facorro  
Kyle Baker  
Luke Imhoff  
Milton Inostroza  
PlatinumThinker  
Yago Riveiro  
fred



# **Part I**

# **Understanding ERTS**



## Chapter 1

# Introducing the Erlang Runtime System

The Erlang RunTime System (ERTS) is a complex system with many interdependent components. It is written in a very portable way so that it can run on anything from a gum stick computer to the largest multicore system with terabytes of memory. In order to be able to optimize the performance of such a system for your application, you need to not only know your application, but you also need to have a thorough understanding of ERTS itself.

## 1.1. ERTS and the Erlang Runtime System

There is a difference between any Erlang Runtime System and a specific implementation of an Erlang Runtime System. "Erlang/OTP" by Ericsson is the de facto standard implementation of Erlang and the Erlang Runtime System. In this book I will refer to this implementation as *ERTS* or spelled out *Erlang RunTime System* with a capital T. (See [ERTS](#) for a definition of OTP).

There is no official definition of what an Erlang Runtime System is, or what an Erlang Virtual Machine is. You could sort of imagine what such an ideal Platonic system would look like by taking ERTS and removing all the implementation specific details. This is unfortunately a circular definition, since you need to know the general definition to be able to identify an implementation specific detail. In the Erlang world we are usually too pragmatic to worry about this.

We will try to use the term *Erlang Runtime System* to refer to the general idea of any Erlang Runtime System as opposed to the specific implementation by Ericsson which we call the Erlang RunTime System or usually just ERTS.

**Note** This book is mostly a book about ERTS in particular and only to a small extent about any general Erlang Runtime System. If you assume that we talk about the Ericsson implementation unless it is clearly stated that we are talking about a general principle you will probably be right.

## 1.2. How to read this book

In [Running ERTS](#) of this book we will look at how to tune the runtime system for your application and how to profile and debug your application and the runtime system. In order to really know how to tune the system you also need to know the system. In [Understanding ERTS](#) of this book you will get a deep understanding of how the runtime system works.

The following chapters of [Understanding ERTS](#) will go over each component of the system by itself. You should be able to read any one of these chapters without having a full understanding of how the other components are implemented, but you will need a basic understanding of what each component is. The rest of this introductory chapter should give you enough basic understanding and vocabulary to be able to jump between the rest of the chapters in part one in any order you like.

However, if you have the time, read the book in order the first time. Words that are specific to Erlang and ERTS or used in a specific way in this book are usually explained at their first occurrence. Then, when you know the vocabulary, you can come back and use Part I as a reference whenever you have a problem with a particular component.

## 1.3. ERTS

In this there is a basic overview of the main components of ERTS and some vocabulary needed to understand the more detailed descriptions of each component in the following chapters.

### 1.3.1. The Erlang Node (ERTS)

When you start an Elixir or Erlang application or system, what you really start is an Erlang node. The node runs the Erlang RunTime System and the virtual machine BEAM. (or possibly another implementation of Erlang (see [Other Erlang](#)

[Implementations](#))).

Your application code will run in an Erlang node, and all the layers of the node will affect the performance of your application. We will look at the stack of layers that makes up a node. This will help you understand your options for running your system in different environments.

In OO terminology one could say that an Erlang node is an object of the Erlang Runtime System class. The equivalent in the Java world is a JVM instance.

All execution of Elixir/Erlang code is done within a node. An Erlang node runs in one OS process, and you can have several Erlang nodes running on one machine.

To be completely correct according to the Erlang OTP documentation a node is actually an executing runtime system that has been given a name. That is, if you start Elixir without giving a name through one of the command line switches `--name NAME@HOST` or `--sname NAME` (or `-name` and `-sname` for an Erlang runtime.) you will have a runtime but not a node. In such a system the function `Node.alive?` (or in Erlang `is_alive()`) returns false.

```
$ iex
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
[async-threads:10] [hipe] [kernel-poll:false]

Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iex(1)> Node.alive?
false
iex(2)>
```

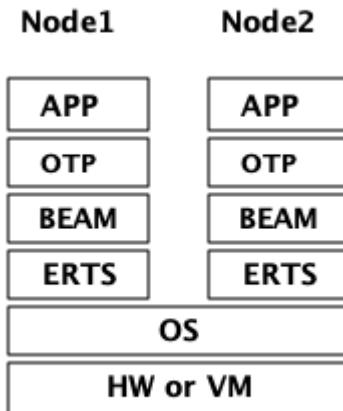
The runtime system itself is not that strict in its use of the terminology. You can ask for the name of the node even if you didn't give it a name. In Elixir you use the function `Node.list` with the argument `:this`, and in Erlang you call `nodes(this)`:

```
iex(2)> Node.list :this
[:nonode@nohost]
iex(3)>
```

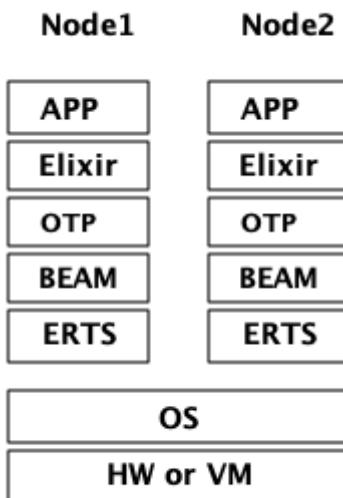
In this book we will use the term node for any running instance of the runtime whether it is given a name or not.

### 1.3.2. Layers in the Execution Environment

Your program (application) will run in one or more nodes, and the performance of your program will depend not only on your application code but also on all the layers below your code in the *ERTS stack*. In [ERTS Stack](#) you can see the ERTS Stack illustrated with two Erlang nodes running on one machine.

*Figure 1. ERTS Stack*

If you are using Elixir there is yet another layer to the stack.

*Figure 2. Elixir Stack*

Let's look at each layer of the stack and see how you can tune them to your application's need.

At the bottom of the stack there is the hardware you are running on. The easiest way to improve the performance of your app is probably to run it on better hardware. You might need to start exploring higher levels of the stack if economical or physical constraints or environmental concerns won't let you upgrade your hardware.

The two most important choices for your hardware is whether it is multicore and whether it is 32-bit or 64-bit. You need different builds of ERTS depending on whether you want to use multicore or not and whether you want to use 32-bit or 64-bit.

The second layer in the stack is the OS level. ERTS runs on most versions of Windows and most POSIX "compliant" operating systems, including Linux, VxWorks, FreeBSD, Solaris, and Mac OS X. Today most of the development of ERTS is done on Linux and OS X, and you can expect the best performance on these platforms. However, Ericsson has been using Solaris internally in many projects and ERTS has been tuned for Solaris for many years. Depending on your use case you might actually get the best performance on a Solaris system. The OS choice is usually not based on performance requirements, but is restricted by other factors. If you are building an embedded application you might be restricted to Raspbian or VxWork, and if you for some reason are building an end user or client application you might have to use Windows. The Windows port of ERTS has so far not had the highest priority and might not be the best choice from a performance or maintenance perspective. If you want to use a 64-bit ERTS

you of course need to have both a 64-bit machine and a 64-bit OS. We will not cover many OS specific questions in this book, and most examples will be assuming that you run on Linux.

The third layer in the stack is the Erlang Runtime System. In our case this will be ERTS. This and the fourth layer, the Erlang Virtual Machine (BEAM), is what this book is all about.

The fifth layer, OTP, supplies the Erlang standard libraries. OTP originally stood for "Open Telecom Platform" and was a number of Erlang libraries supplying building blocks (such as `supervisor`, `gen_server` and `gen_tcp`) for building robust applications (such as telephony exchanges). Early on, the libraries and the meaning of OTP got intermingled with all the other standard libraries shipped with ERTS. Nowadays most people use OTP together with Erlang in "Erlang/OTP" as the name for ERTS and all Erlang libraries shipped by Ericsson. Knowing these standard libraries and how and when to use them can greatly improve the performance of your application. This book will not go into any details of the standard libraries and OTP, there are many other books that cover these aspects.

If you are running an Elixir program the sixth layer provides the Elixir environment and the Elixir libraries.

Finally, the seventh layer (APP) is your application, and any third party libraries you use. The application can use all the functionality provided by the underlying layers. Apart from upgrading your hardware this is probably the place where you most easily can improve your application's performance. In [Tracing](#) there are some hints and some tools that can help you profile and optimize your application. In [Debugging](#) we will look at how to find the cause of crashing applications and how to find bugs in your application.

For information on how to build and run an Erlang node see [\[CH-BuildingERTS\]](#), and read the rest of the book to learn all about the components of an Erlang node.

### 1.3.3. Distribution

One of the key insights by the Erlang language designers was that in order to build a system that works 24/7 you need to be able to handle hardware failure. Therefore you need to distribute your system over at least two physical machines. You do this by starting a node on each machine and then you can connect the nodes to each other so that processes can communicate with each other across the nodes just as if they were running in the same node.

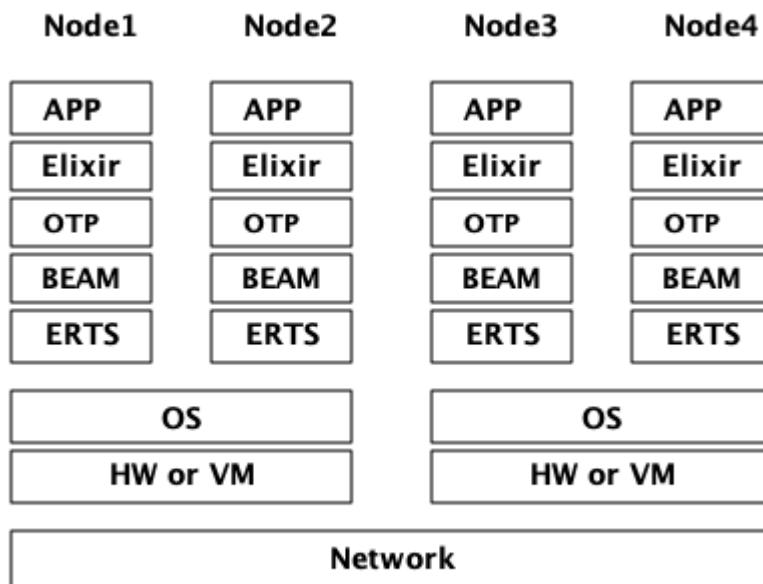


Figure 3. Distributed Applications

### 1.3.4. The Erlang Compiler

The Erlang Compiler is responsible for compiling Erlang source code, from .erl files into virtual machine code for BEAM (the

virtual machine). The compiler itself is written in Erlang and compiled by itself to BEAM code and usually available in a running Erlang node. To bootstrap the runtime system there are a number of precompiled BEAM files, including the compiler, in the bootstrap directory.

For more information about the compiler see [The Compiler](#).

### 1.3.5. The Erlang Virtual Machine: BEAM

BEAM is the Erlang virtual machine used for executing Erlang code, just like the JVM is used for executing Java code. BEAM runs in an Erlang Node.

**BEAM:** The name BEAM originally stood for Bogdan's Erlang Abstract Machine, but nowadays most people refer to it as Björn's Erlang Abstract Machine, after the current maintainer.

Just as ERTS is an implementation of a more general concept of a Erlang Runtime System so is BEAM an implementation of a more general Erlang Virtual Machine (EVM). There is no definition of what constitutes an EVM but BEAM actually has two levels of instructions *Generic Instructions* and *Specific Instructions*. The generic instruction set could be seen as a blueprint for an EVM.

For a full description of BEAM see [The Erlang Virtual Machine: BEAM, Modules and The BEAM File Format \(16p\)](#) and [Generic BEAM Instructions \(25p\)](#).

### 1.3.6. Processes

An Erlang process basically works like an OS process. Each process has its own memory (a mailbox, a heap and a stack) and a process control block (PCB) with information about the process.

All Erlang code execution is done within the context of a process. One Erlang node can have many processes, which can communicate through message passing and signals. Erlang processes can also communicate with processes on other Erlang nodes as long as the nodes are connected.

To learn more about processes and the PCB see [Processes](#).

### 1.3.7. Scheduling

The Scheduler is responsible for choosing the Erlang process to execute. Basically the scheduler keeps two queues, a *ready queue* of processes ready to run, and a *waiting queue* of processes waiting to receive a message. When a process in the waiting queue receives a message or gets a time out it is moved to the ready queue.

The scheduler picks the first process from the ready queue and hands it to BEAM for execution of one *time slice*. BEAM preempts the running process when the time slice is used up and adds the process to the end of the ready queue. If the process is blocked in a receive before the time slice is used up, it gets added to the waiting queue instead.

Erlang is concurrent by nature, that is, each process is conceptually running at the same time as all other processes, but in reality there is just one process running in the VM. On a multicore machine Erlang actually runs more than one scheduler, usually one per physical core, each having their own queues. This way Erlang achieves true parallelism. To utilize more than one core ERTS has to be built (see [\[CH-BuildingERTS\]](#)) in *SMP* mode. SMP stands for *Symmetric MultiProcessing*, that is, the ability to execute a processes on any one of multiple CPUs.

In reality the picture is more complicated with priorities among processes and the waiting queue is implemented through a timing wheel. All this and more is described in detail in [Scheduling](#).

### 1.3.8. The Erlang Tag Scheme

Erlang is a dynamically typed language, and the runtime system needs a way to keep track of the type of each data object. This is done with a tagging scheme. Each data object or pointer to a data object also has a tag with information about the data type of the object.

Basically some bits of a pointer are reserved for the tag, and the emulator can then determine the type of the object by looking at the bit pattern of the tag.

These tags are used for pattern matching and for type test and for primitive operations as well as by the garbage collector.

The complete tagging scheme is described in [The Erlang Type System and Tags](#).

### 1.3.9. Memory Handling

Erlang uses automatic memory management and the programmer does not have to worry about memory allocation and deallocation. Each process has a heap and a stack which both can grow, and shrink, as needed.

When a process runs out of heap space, the VM will first try to reclaim free heap space through garbage collection. The garbage collector will then go through the process stack and heap and copy live data to a new heap while throwing away all the data that is dead. If there still isn't enough heap space, a new larger heap will be allocated and the live data is moved there.

The details of the current generational copying garbage collector, including the handling of reference counted binaries can be found in [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#).

In a system which uses HiPE compiled native code, each process actually has two stacks, a BEAM stack and a native stack, the details can be found in [Native Code](#).

### 1.3.10. The Interpreter and the Command Line Interface

When you start an Erlang node with erl you get a command prompt. This is the *Erlang read eval print loop* (REPL) or the *command line interface* (CLI) or simply the *Erlang shell*.

You can actually type in Erlang code and execute it directly from the shell. In this case the code is not compiled to BEAM code and executed by the BEAM, instead the code is parsed and interpreted by the Erlang interpreter. In general the interpreted code behaves exactly as compiled code, but there are a few subtle differences, these differences and all other aspects of the shell are explained in [Operation and Maintenance](#).

## 1.4. Other Erlang Implementations

This book is mainly concerned with the "standard" Erlang implementation by Ericsson/OTP called ERTS, but there are a few other implementations available and in this section we will look at some of them briefly.

### 1.4.1. Erlang on Xen

Erlang on Xen (<http://erlangonxen.org>) is an Erlang implementation running directly on server hardware with no OS layer in between, only a thin Xen client.

Ling, the virtual machine of Erlang on Xen is almost 100% binary compatible with BEAM. In [xref:the\\_eox\\_stack](#) you can see how the Erlang on Xen implementation of the Erlang Solution Stack differs from the ERTS Stack. The thing to note here is that there is no operating system in the Erlang on Xen stack.

Since Ling implements the generic instruction set of BEAM, it can reuse the BEAM compiler from the OTP layer to compile Erlang to Ling.

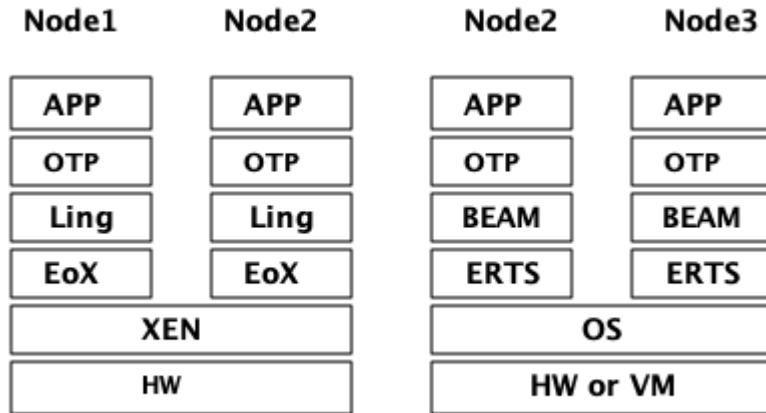


Figure 4. Erlang On Xen

### 1.4.2. Erjang

Erjang (<http://www.erjang.org>) is an Erlang implementation which runs on the JVM. It loads .beam files and recompiles the code to Java .class files. Erjang is almost 100% binary compatible with (generic) BEAM.

In `xref:the_erjang_stack` you can see how the Erjang implementation of the Erlang Solution Stack differs from the ERTS Stack. The thing to note here is that JVM has replaced BEAM as the virtual machine and that Erjang provides the services of ERTS by implementing them in Java on top of the VM.

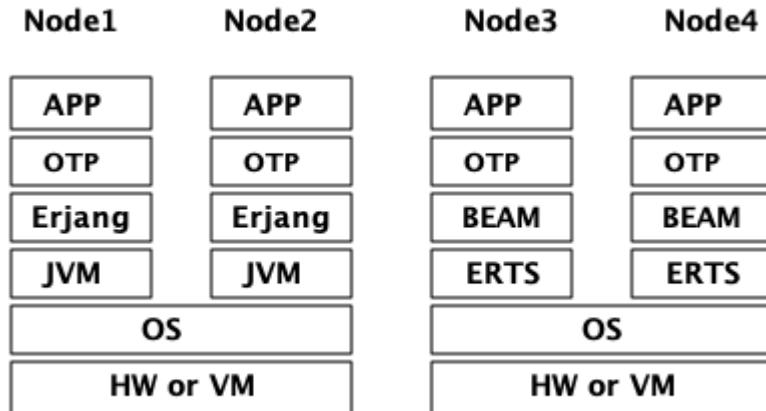


Figure 5. Erlang on the JVM

Now that you have a basic understanding of all the major pieces of ERTS, and the necessary vocabulary you can dive into the details of each component. If you are eager to understand a certain component, you can jump directly to that chapter. Or if you are really eager to find a solution to a specific problem you could jump to the right chapter in [Running ERTS](#), and try the different methods to tune, tweak, or debug your system.



## Chapter 2

# The Compiler

This book will not cover the programming language Erlang, but since the goal of the ERTS is to run Erlang code you will need to know how to compile Erlang code. In this chapter we will cover the compiler options needed to generate readable beam code and how to add debug information to the generated beam file. At the end of the chapter there is also a section on the Elixir compiler.

For those readers interested in compiling their own favorite language to ERTS this chapter will also contain detailed information about the different intermediate formats in the compiler and how to plug your compiler into the beam compiler backend. I will also present parse transforms and give examples of how to use them to tweak the Erlang language.

## 2.1. Compiling Erlang

Erlang is compiled from source code modules in .erl files to fat binary .beam files.

The compiler can be run from the OS shell with the erlc command:

```
> erlc foo.erl
```

Alternatively the compiler can be invoked from the Erlang shell with the default shell command c or by calling compile:file/{1,2}

```
1> c(foo).
```

or

```
1> compile:file(foo).
```

The optional second argument to compile:file is a list of compiler options. A full list of the options can be found in the documentation of the compile module: see <http://www.erlang.org/doc/man/compile.html>.

Normally the compiler will compile Erlang source code from a .erl file and write the resulting binary beam code to a .beam file. You can also get the resulting binary back as an Erlang term by giving the option binary to the compiler. This option has then been overloaded to mean return any intermediate format as a term instead of writing to a file. If you for example want the compiler to return Core Erlang code you can give the options [core, binary].

The compiler is made up of a number of passes as illustrated in [Compiler Passes](#).

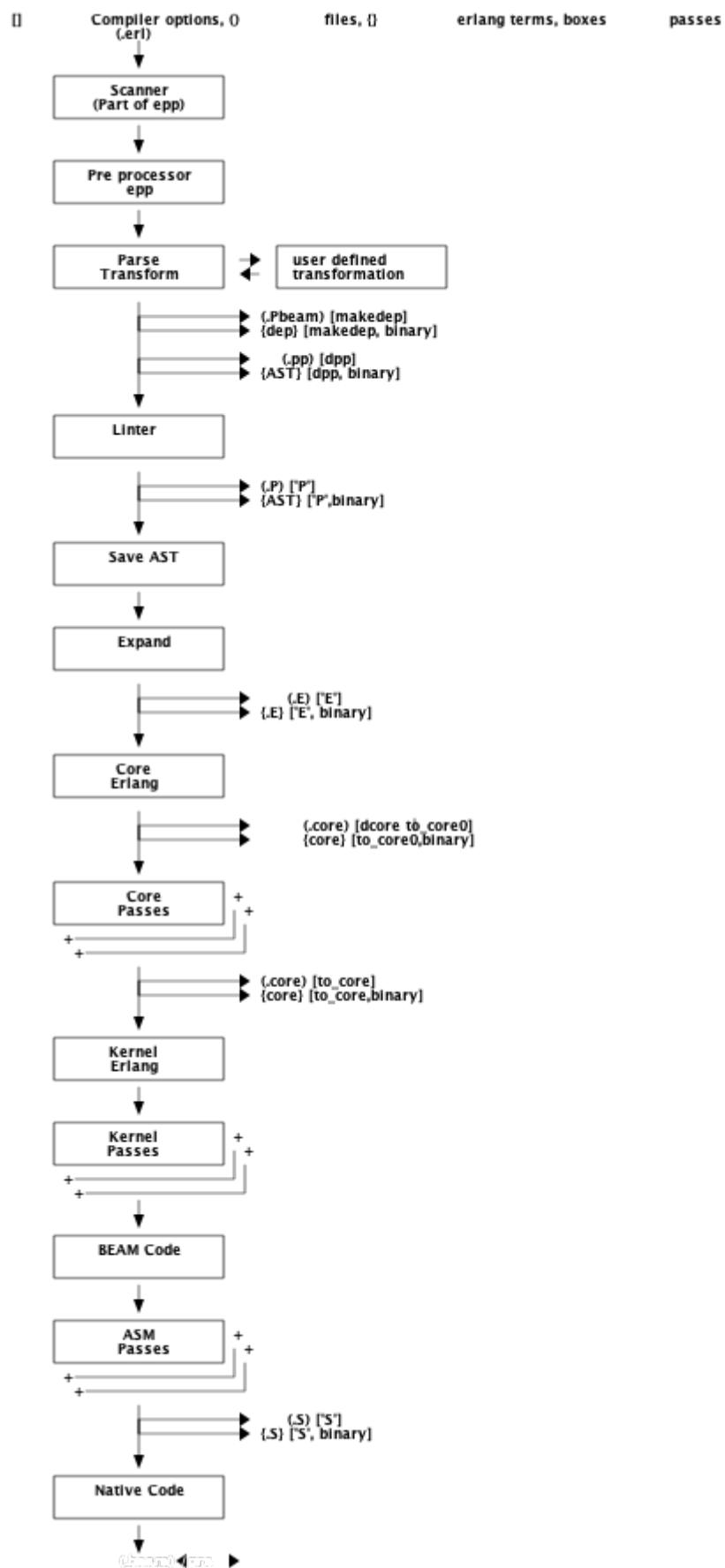


Figure 6. Compiler Passes

If you want to see a complete and up to date list of compiler passes you can run the function `compile:options/0` in an Erlang shell.

The definitive source for information about the compiler is of course the source: [compile.erl](https://github.com/erlang/otp/blob/maint/lib/compiler/src/compile.erl) [https://github.com/erlang/otp/blob/maint/lib/compiler/src/compile.erl]

## 2.2. Generating Intermediate Output

Looking at the code produced by the compiler is a great help in trying to understand how the virtual machine works. Fortunately, the compiler can show us the intermediate code after each compiler pass and the final beam code.

Let us try out our newfound knowledge to look at the generated code.

```
1> compile:options().
dpp - Generate .pp file
'P' - Generate .P source listing file
```

...

```
'E' - Generate .E source listing file
```

...

```
'S' - Generate .S file
```

Let us try with a small example program "world.erl":

```
-module(world).
-export([hello/0]).

-include("world.hrl").

hello() -> ?GREETING.
```

And the include file "world.hrl"

```
-define(GREETING, "hello world").
```

If you now compile this with the 'P' option to get the parsed file you get a file "world.P":

```
2> c(world, ['P']).
** Warning: No object file created - nothing loaded **
ok
```

In the resulting .P file you can see a pretty printed version of the code after the preprocessor (and parse transformation) has been applied:

```
-file("world.erl", 1).

-module(world).

-export([hello/0]).

-file("world.hrl", 1).

-file("world.erl", 4).

hello() ->
    "hello world".
```

To see how the code looks after all source code transformations are done, you can compile the code with the 'E'-flag.

```
3> c(world, ['E']).
** Warning: No object file created - nothing loaded **
ok
```

This gives us an .E file, in this case all compiler directives have been removed and the built in functions module\_info/{1,2} have been added to the source:

```
-vsn("\002").

-file("world.erl", 1).

-file("world.hrl", 1).

-file("world.erl", 5).

hello() ->
    "hello world".

module_info() ->
    erlang:get_module_info(world).

module_info(X) ->
    erlang:get_module_info(world, X).
```

We will make use of the 'P' and 'E' options when we look at parse transforms in [Compiler Pass: Parse Transformations](#), but first we will take a look at an "assembler" view of generated BEAM code. By giving the option 'S' to the compiler you get a .S file with Erlang terms for each BEAM instruction in the code.

```
3> c(world, ['S']).
** Warning: No object file created - nothing loaded **
ok
```

The file world.S should look like this:

```

{module, world}. %% version = 0

{exports, [{hello,0},{module_info,0},{module_info,1}]}.

{attributes, []}.

{labels, 7}.

{function, hello, 0, 2}.
{label,1}.
{line,[{location,"world.erl",6}]}.
{func_info,{atom,world},{atom,hello},0}.
{label,2}.
{move,{literal,"hello world"},{x,0}}.
return.

{function, module_info, 0, 4}.
{label,3}.
{line,[]}.
{func_info,{atom,world},{atom,module_info},0}.
{label,4}.
{move,{atom,world},{x,0}}.
{line,[]}.
{call_ext_only,1,{extfunc,erlang,get_module_info,1}}.

{function, module_info, 1, 6}.
{label,5}.
{line,[]}.
{func_info,{atom,world},{atom,module_info},1}.
{label,6}.
{move,{x,0},{x,1}}.
{move,{atom,world},{x,0}}.
{line,[]}.
{call_ext_only,2,{extfunc,erlang,get_module_info,2}}.

```

Since this is a file with dot (".") separated Erlang terms, you can read the file back into the Erlang shell with:

```
{ok, BEAM_Code} = file:consult("world.S").
```

The assembler code mostly follows the layout of the original source code.

The first instruction defines the module name of the code. The version mentioned in the comment (%% version = 0) is the version of the beam opcode format (as given by beam\_opcodes:format\_number/0).

Then comes a list of exports and any compiler attributes (none in this example) much like in any Erlang source module.

The first real beam-like instruction is {labels, 7} which tells the VM the number of labels in the code to make it possible to allocate room for all labels in one pass over the code.

After that there is the actual code for each function. The first instruction gives us the function name, the arity and the entry point as

a label number.

You can use the 'S' option with great effect to help you understand how the BEAM works, and we will use it like that in later chapters. It is also invaluable if you develop your own language that you compile to the BEAM through Core Erlang, to see the generated code.

## 2.3. Compiler Passes

In the following sections we will go through most of the compiler passes shown in [Compiler Passes](#). For a language designer targeting the BEAM this is interesting since it will show you what you can accomplish with the different approaches: macros, parse transforms, core erlang, and BEAM code, and how they depend on each other.

When tuning Erlang code, it is good to know what optimizations are applied when, and how you can look at generated code before and after optimizations.

### 2.3.1. Compiler Pass: The Erlang Preprocessor (epp)

The compilation starts with a combined tokenizer (or scanner) and preprocessor. That is, the preprocessor drives the tokenizer. This means that macros are expanded as tokens, so it is not a pure string replacement (as for example m4 or cpp). You can not use Erlang macros to define your own syntax, a macro will expand as a separate token from its surrounding characters. You can not concatenate a macro and a character to a token:

```
-define(plus,+).
t(A,B) -> A?plus+B.
```

This will expand to

```
t(A,B) -> A + + B.
```

and not

```
t(A,B) -> A ++ B.
```

On the other hand since macro expansion is done on the token level, you do not need to have a valid Erlang term in the right hand side of the macro, as long as you use it in a way that gives you a valid term. E.g.:

```
-define(p,o, o]).
t() -> [f,?p.
```

There are few real usages for this other than to win the obfuscated Erlang code contest. The main point to remember is that you can not really use the Erlang preprocessor to define a language with a syntax that differs from Erlang. Fortunately there are other ways to do this, as you shall see later.

### 2.3.2. Compiler Pass: Parse Transformations

The easiest way to tweak the Erlang language is through Parse Transformations (or parse transforms). Parse Transformations come with all sorts of warnings, like this note in the OTP documentation:



Programmers are strongly advised not to engage in parse transformations and no support is offered for problems encountered.

When you use a parse transform you are basically writing an extra pass in the compiler and that can if you are not careful lead to very unexpected results. But to use a parse transform you have to declare the usage in the module using it, and it will be local to that module, so as far as compiler tweaks goes this one is quite safe.

The biggest problem with parse transforms as I see it is that you are inventing your own syntax, and it will make it more difficult for anyone else reading your code. At least until your parse transform has become as popular and widely used as e.g. QLC.

OK, so you know you shouldn't use it, but if you have to, here is what you need to know. A parse transforms is a function that works on the abstract syntax tree (AST) (see <http://www.erlang.org/doc/apps/erts/absform.html>). The compiler does preprocessing, tokenization and parsing and then it will call the parse transform function with the AST and expects to get back a new AST.

This means that you can't change the Erlang syntax fundamentally, but you can change the semantics. Lets say for example that you for some reason would like to write json code directly in your Erlang code, then you are in luck since the tokens of json and of Erlang are basically the same. Also, since the Erlang compiler does most of its sanity checks in the linter pass which follows the parse transform pass, you can allow an AST which does not represent valid Erlang.

To write a parse transform you need to write an Erlang module (lets call it *p*) which exports the function `parse_transform/2`. This function is called by the compiler during the parse transform pass if the module being compiled (lets call it *m*) contains the compiler option `{parse_transform, p}`. The arguments to the function is the AST of the module *m* and the compiler options given to the call to the compiler.

 Note that you will not get any compiler options given in the file, this is a bit of a nuisance since you can't give options to the parse transform from the code.

The compiler does not expand compiler options until the *expand* pass which occurs after the parse transform pass.

The documentation of the abstract format is somewhat dense and it is quite hard to get a grip on the abstract format by reading the documentation. I encourage you to use the `syntax_tools` and especially `erl_syntax_lib` for any serious work on the AST.

Here we will develop a simple parse transform just to get an understanding of the AST. Therefore we will work directly on the AST and use the old reliable `io:format` approach instead of `syntax_tools`.

First we create an example of what we would like to be able to compile `json_test.erl`:

```
-module(json_test).
-compile({parse_transform, json_parser}).
-export([test/1]).

test(V) ->
  <<{{
    "name" : "Jack (\\"Bee\\") Nimble",
    "format": {
      "type"      : "rect",
      "widths"    : [1920,1600],
      "height"   : (-1080),
      "interlace" : false,
      "frame rate": V
    }
  }}>>.
```

Then we create a minimal parse transform module `json_parser.erl`:

```
-module(json_parser).
-export([parse_transform/2]).

parse_transform(AST, _Options) ->
  io:format("~p~n", [AST]),
  AST.
```

This identity parse transform returns an unchanged AST but it also prints it out so that you can see what an AST looks like.

```
> c(json_parser).
{ok,json_parser}
2> c(json_test).
[{attribute,1,file,{"./json_test.erl",1}},
 {attribute,1,module,json_test},
 {attribute,3,export,[{test,1}]},
 {function,5,test,1,
  [{clause,5,
    [{var,5,'V'}],
    [],
    [{bin,6,
      [{bin_element,6,
        {tuple,6,
         [{tuple,6,
           [{remote,7,{string,7,"name"},{string,7,"Jack (\\"Bee\\") Nimble"}},
            {remote,8,
             {string,8,"format"}},
            {tuple,8,
             [{remote,9,{string,9,"type"},{string,9,"rect"}},
              {remote,10,
               {string,10,"widths"}},
              {cons,10,
               {integer,10,1920},
               {cons,10,{integer,10,1600},{nil,10}}}}},
            {remote,11,{string,11,"height"},{op,11,'-',{integer,11,1080}}},
            {remote,12,{string,12,"interlace"},{atom,12,false}},
            {remote,13,{string,13,"frame rate"},{var,13,'V'}}]}]}]}],
    default,default}]}]},],
{eof,16}]
./json_test.erl:7: illegal expression
./json_test.erl:8: illegal expression
./json_test.erl:5: Warning: variable 'V' is unused
error
```

The compilation of json\_test fails since the module contains invalid Erlang syntax, but you get to see what the AST looks like. Now we can just write some functions to traverse the AST and rewrite the json code into Erlang code.<sup>[1]</sup>

```
-module(json_parser).
-export([parse_transform/2]).

parse_transform(AST, _Options) ->
  json(AST, []).
```

```

-define(FUNCTION(Clauses), {function, Label, Name, Arity, Clauses}).

%% We are only interested in code inside functions.
json([?FUNCTION(Clauses) | Elements], Res) ->
    json(Elements, [?FUNCTION(json_clauses(Clauses)) | Res]);
json([Other|Elements], Res) -> json(Elements, [Other | Res]);
json([], Res) -> lists:reverse(Res).

%% We are interested in the code in the body of a function.
json_clauses([{clause, CLine, A1, A2, Code} | Clauses]) ->
    [{clause, CLine, A1, A2, json_code(Code)} | json_clauses(Clauses)];
json_clauses([]) -> [].

-define(JSON(Json), {bin, _, [{bin_element
        , _}, {tuple, _, [Json]}]
        , _]}).

%% We look for: <<"json">> = Json-Term
json_code([]) -> [];
json_code([?JSON(Json)|MoreCode]) -> [parse_json(Json) | json_code(MoreCode)];
json_code(Code) -> Code.

%% Json Object -> []
parse_json({tuple, Line, []}) -> {cons, Line, {tuple, Line, []}};
parse_json({tuple, Line, Fields}) -> parse_json_fields(Fields, Line);
%% Json Array -> List
parse_json({cons, Line, Head, Tail}) -> {cons, Line, parse_json(Head),
                                             parse_json(Tail)};
parse_json({nil, Line}) -> {nil, Line};
%% Json String -> <<String>>
parse_json({string, Line, String}) -> str_to_bin(String, Line);
%% Json Integer -> Intger
parse_json({integer, Line, Integer}) -> {integer, Line, Integer};
%% Json Float -> Float
parse_json({float, Line, Float}) -> {float, Line, Float};
%% Json Constant -> true | false | null
parse_json({atom, Line, true}) -> {atom, Line, true};
parse_json({atom, Line, false}) -> {atom, Line, false};
parse_json({atom, Line, null}) -> {atom, Line, null};

%% Variables, should contain Erlang encoded Json
parse_json({var, Line, Var}) -> {var, Line, Var};
%% Json Negative Integer or Float
parse_json({op, Line, '-', {Type, _, N}}) when Type ==:= integer
                                              ; Type ==:= float ->
    {Type, Line, -N}.
%% parse_json(Code) -> io:format("Code: ~p~n",[Code]), Code.

-define(FIELD(Label, Code), {remote, L, {string, _, Label}, Code}).

```

```

parse_json_fields([], L) -> {nil, L};
%% Label : Json-Term --> [{<<Label>>, Term} | Rest]
parse_json_fields([?FIELD(Label, Code) | Rest], _) ->
  cons(tuple(str_to_bin(Label, L), parse_json(Code), L)
    , parse_json_fields(Rest, L)
    , L).

tuple(E1, E2, Line) -> {tuple, Line, [E1, E2]}.
cons(Head, Tail, Line) -> {cons, Line, Head, Tail}.

str_to_bin(String, Line) ->
  {bin
  , Line
  , [{bin_element
      , Line
      , {string, Line, String}
      , default
      , default
    }
  ]
}.

```

And now we can compile json\_test without errors:

```

1> c(json_parser).
{ok,json_parser}
2> c(json_test).
{ok,json_test}
3> json_test:test(42).
[{<<"name">>,<<"Jack (\\"Bee\\") Nimble">>},
 {<<"format">>,
  [{<<"type">>,<<"rect">>},
   {<<"widths">>,[1920,1600]},
   {<<"height">>,-1080},
   {<<"interlace">>,false},
   {<<"frame rate">>,42}]}]

```

The AST generated by parse\_transform/2 must correspond to valid Erlang code unless you apply several parse transforms, which is possible. The validity of the code is checked by the following compiler pass.

### 2.3.3. Compiler Pass: Linter

The linter (erl\_lint.erl) generates warnings for syntactically correct but otherwise bad code, like "export\_all flag enabled".

### 2.3.4. Compiler Pass: Save AST

In order to enable debugging of a module, you can "debug compile" the module, that is to pass the option debug\_info to the compiler. The abstract syntax tree will then be saved by the "Save AST" until the end of the compilation, where it will be written to the .beam file.

It is important to note that the code is saved before any optimisations are applied, so if there is a bug in an optimisation pass in the compiler and you run code in the debugger you will get a different behavior. If you are implementing your own compiler

optimisations this can trick you up badly.

### 2.3.5. Compiler Pass: Expand

In the expand phase source erlang constructs, such as records, are expanded to lower level erlang constructs. Compiler options, "-compile(...)", are also *expanded* to meta data.

### 2.3.6. Compiler Pass: Core Erlang

Core Erlang is a strict functional language suitable for compiler optimizations. It makes code transformations easier by reducing the number of ways to express the same operation. One way it does this is by introducing *let* and *letrec* expressions to make scoping more explicit.

Core Erlang is the best target for a language you want to run in ERTS. It changes very seldom and it contains all aspects of Erlang in a clean way. If you target the beam instruction set directly you will have to deal with much more details, and that instruction set usually changes slightly between each major release of ERTS. If you on the other hand target Erlang directly you will be more restricted in what you can describe, and you will also have to deal with more details, since Core Erlang is a cleaner language.

To compile an Erlang file to core you can give the option "to\_core", note though that this writes the Erlang core program to a file with the ".core" extension. To compile an Erlang core program from a ".core" file you can give the option "from\_core" to the compiler.

```
1> c(world, to_core).
** Warning: No object file created - nothing loaded **
ok
2> c(world, from_core).
{ok,world}
```

Note that the .core files are text files written in the human readable core format. To get the core program as an Erlang term you can add the binary option to the compilation.

### 2.3.7. Compiler Pass: Kernel Erlang

Kernel Erlang is a flat version of Core Erlang with a few differences. For example, each variable is unique and the scope is a whole function. Pattern matching is compiled to more primitive operations.

### 2.3.8. Compiler Pass: BEAM Code

The last step of a normal compilation is the external beam code format. Some low level optimizations such as dead code elimination and peep hole optimisations are done on this level.

The BEAM code is described in detail in [Generic BEAM Instructions \(25p\)](#) and [BEAM Instructions](#)

### 2.3.9. Compiler Pass: Native Code

If you add the flag native to the compilation, and you have a HiPE enabled runtime system, then the compiler will generate native code for your module and store the native code along with the beam code in the .beam. file.

## 2.4. Other Compiler Tools

There are a number of tools available to help you work with code generation and code manipulation. These tools are written in Erlang and not really part of the runtime system but they are very nice to know about if you are implementing another language on top of the BEAM.

In this section we will cover three of the most useful code tools: the lexer — Leex, the parser generator — Yecc, and a general set

of functions to manipulate abstract forms — Syntax Tools.

## 2.4.1. Leex

Leex is the Erlang lexer generator. The lexer generator takes a description of a DFA from a definitions file xrl and produces an Erlang program that matches tokens described by the DFA.

The details of how to write a DFA definition for a tokenizer is beyond the scope of this book. For a thorough explanation I recommend the "Dragon book" (Compiler ... by Aho, Sethi and Ullman). Other good resources are the man and info entry for "flex" the lexer program that inspired leex, and the leex documentation itself. If you have info and flex installed you can read the full manual by typing:

```
> info flex
```

The online Erlang documentation also has the leex manual (see [yecc.html](#) [<http://erlang.org/doc/man/yecc.html>]).

We can use the lexer generator to create an Erlang program which recognizes JSON tokens. By looking at the JSON definition <http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf> we can see that there are only a handful of tokens that we need to handle.

```
Definitions.

Digit      = [0-9]
Digit1to9   = [1-9]
HexDigit    = [0-9a-f]
UnescapedChar = [\\"\\]
EscapedChar = (\\\\)|(\\"")|(\\\b)|(\\\f)|(\\\n)|(\\\r)|(\\\t)|(\\\/)
Unicode    = (\\\u{HexDigit}{HexDigit}{HexDigit}{HexDigit})
Quote      = [""]
Delim      = [\[\]:,{}]
Space       = [\n\s\t\r]

Rules.

{Quote}{Quote} : {token, {string, TokenLine, ""}}.
{Quote}({EscapedChar}|{UnescapedChar})|({Unicode})+{Quote} :
  {token, {string, TokenLine, drop_quotes(TokenChars)}}.

null : {token, {null, TokenLine}}.
true : {token, {true, TokenLine}}.
false : {token, {false, TokenLine}}.

{Delim} : {token, {list_to_atom(TokenChars), TokenLine}}.

{Space} : skip_token.

-{Digit1to9}+{Digit}*.{Digit}+((E|e)(\+|\-)?{Digit}+) :
  {token, {number, TokenLine, list_to_float(TokenChars)}}.
-{Digit1to9}+{Digit}* :
  {token, {number, TokenLine, list_to_integer(TokenChars)+0.0}}.

Erlang code.
-export([t/0]).
```

```

drop_quotes([$" | QuotedString]) -> literal(lists:droplast(QuotedString)).
literal([$\\,$" | Rest]) ->
  [$"|literal(Rest)];
literal([$\\,$\\ | Rest]) ->
  [$\\|literal(Rest)];
literal([$\\,$/ | Rest]) ->
  [$/|literal(Rest)];
literal([$\\,$b | Rest]) ->
  [$\b|literal(Rest)];
literal([$\\,$f | Rest]) ->
  [$\f|literal(Rest)];
literal([$\\,$n | Rest]) ->
  [$\n|literal(Rest)];
literal([$\\,$r | Rest]) ->
  [$\r|literal(Rest)];
literal([$\\,$t | Rest]) ->
  [$\t|literal(Rest)];
literal([$\\,$u,D0,D1,D2,D3|Rest]) ->
  Char = list_to_integer([D0,D1,D2,D3],16),
  [Char|literal(Rest)];
literal([C|Rest]) ->
  [C|literal(Rest)];
literal([]) ->[].

t() ->
{ok,
 [{'{',1},
  {string,2,"no"},
  {':',2},
  {number,2,1.0},
  {'}',3}
 ],
 4}.

```

By using the Leex compiler we can compile this DFA to Erlang code, and by giving the option `dfa_graph` we also produce a dot-file which can be viewed with e.g. Graphviz.

```

1> leex:file(json_tokens, [dfa_graph]).
{ok, "./json_tokens.erl"}
2>

```

You can view the DFA graph using for example dotty.

```
> dotty json_tokens.dot
```

[json tokens] | ..../code/compiler\_chapter/json\_tokens.png

We can try our tokenizer on an example json file (test.json).

```
{
  "no" : 1,
  "name" : "Jack \"Bee\" Nimble",
  "escapes" : "\b\n\r\t\f\\\"",
  "format": {
    "type"      : "rect",
    "widths"    : [1920,1600],
    "height"   : -1080,
    "interlace" : false,
    "unicode"   : "\u002f",
    "frame rate": 4.5
  }
}
```

First we need to compile our tokenizer, then we read the file and convert it to a string. Finally we can use the string/1 function that leex generates to tokenize the test file.

```
2> c(json_tokens).
{ok,json_tokens}.
3> f(File), f(L), {ok, File} = file:read_file("test.json"), L = binary_to_list(File), ok.
ok
4> f(Tokens), {ok, Tokens,_} = json_tokens:string(L), hd(Tokens).
{'{',1}
5>
```

The shell function f/1 tells the shell to forget a variable binding. This is useful if you want to try a command that binds a variable multiple times, for example as you are writing the lexer and want to try it out after each rewrite. We will look at the shell commands in detail in the later chapter.

Armed with a tokenizer for JSON we can now write a json parser using the parser generator Yecc.

## 2.4.2. Yecc

Yecc is a parser generator for Erlang. The name comes from Yacc (Yet another compiler compiler) the canonical parser generator for C.

Now that we have a lexer for JSON terms we can write a parser using yecc.

**Nonterminals** value values object array pair pairs.

**Terminals** number string true false null '[' ']' '{' '}' ',', ':';.

**Rootsymbol** value.

```

value -> object : '$1'.
value -> array : '$1'.
value -> number : get_val('$1').
value -> string : get_val('$1').
value -> 'true' : get_val('$1').
value -> 'null' : get_val('$1').
value -> 'false' : get_val('$1').

object -> '{' '}' : #{}.
object -> '{' pairs '}' : '$2'.

pairs -> pair : '$1'.
pairs -> pair ',' pairs : maps:merge('$1', '$3').

pair -> string ':' value : #{ get_val('$1') => '$3' }.

array -> '[' ']' : {}.
array -> '[' values ']' : list_to_tuple('$2').

values -> value : [ '$1' ].
values -> value ',' values : [ '$1' | '$3' ].
```

Erlang code.

```

get_val({_, _, Val}) -> Val;
get_val({Val, _}) -> Val.
```

Then we can use yecc to generate an Erlang program that implements the parser, and call the parse/1 function provided with the tokens generated by the tokenizer as an argument.

```

5> yecc:file(yecc_json_parser), c(yecc_json_parser).
{ok, yexx_json_parser}
6> f(Json), {ok, Json} = yecc_json_parser:parse(Tokens).
{ok, #{"escapes" => "\b\n\r\t\f///",
      "format" => #{"frame_rate" => 4.5,
                    "height" => -1080.0,
                    "interlace" => false,
                    "type" => "rect",
                    "unicode" => "/",
                    "widths" => {1920.0, 1.6e3}},
       "name" => "Jack \"Bee\" Nimble",
       "no" => 1.0}}
```

The tools Leex and Yecc are nice when you want to compile your own complete language to the Erlang virtual machine. By

combining them with Syntax tools and specifically Merl you can manipulate the Erlang Abstract Syntax tree, either to generate Erlang code or to change the behaviour of Erlang code.

## 2.5. Syntax Tools and Merl

Syntax Tools is a set of libraries for manipulating the internal representation of Erlang's Abstract Syntax Trees (ASTs).

The syntax tools applications also includes the tool Merl since Erlang 18.0. With Merl you can very easily manipulate the syntax tree and write parse transforms in Erlang code.

You can find the documentation for Syntax Tools on the Erlang.org site: [http://erlang.org/doc/apps/syntax\\_tools/chapter.html](http://erlang.org/doc/apps/syntax_tools/chapter.html) [[http://erlang.org/doc/apps/syntax\\_tools/chapter.html](http://erlang.org/doc/apps/syntax_tools/chapter.html)].

## 2.6. Compiling Elixir

Another approach to writing your own language on top of the Beam is to use the meta programming tools in Elixir. Elixir compiles to Beam code through the Erlang abstract syntax tree.

With Elixir's defmacro you can define your own Domain Specific Language, directly in Elixir.

[1] The translation here is done in accordance with [EEP 18](http://www.erlang.org/eeps/eep-0018.html) [<http://www.erlang.org/eeps/eep-0018.html>] (Erlang Enhancement Proposal 18: "JSON bifs")

# Chapter 3

# Processes

The concept of lightweight processes is the essence of Erlang and the BEAM; they are what makes BEAM stand out from other virtual machines. In order to understand how the BEAM (and Erlang and Elixir) works you need to know the details of how processes work, which will help you understand the central concept of the BEAM, including what is easy and cheap for a process and what is hard and expensive.

Almost everything in the BEAM is connected to the concept of processes and in this chapter we will learn more about these connections. We will expand on what we learned in the introduction and take a deeper look at concepts such as memory management, message passing, and in particular scheduling.

An Erlang process is very similar to an OS process. It has its own address space, it can communicate with other processes through signals and messages, and the execution is controlled by a preemptive scheduler.

When you have a performance problem in an Erlang or Elixir system the problem is very often stemming from a problem within a particular process or from an imbalance between processes. There are of course other common problems such as bad algorithms or memory problems which we will look at in other chapters. Still, being able to pinpoint the process which is causing the problem is always important, therefore we will look at the tools available in the Erlang RunTime System for process inspection.

We will introduce the tools throughout the chapter as we go through how a process and the scheduler works, and then we will bring all tools together for an exercise at the end.

## 3.1. What is a Process?

A process is an isolated entity where code execution occurs. A process protects your system from errors in your code by isolating the effect of the error to the process executing the faulty code.

The runtime comes with a number of tools for inspecting processes to help us find bottlenecks, problems and overuse of resources. These tools will help you identify and inspect problematic processes.

### 3.1.1. Listing Processes from the Shell

Let us dive right in and look at which processes we have in a running system. The easiest way to do that is to just start an Erlang shell and issue the shell command `i()`. In Elixir you can call the function in the `:shell_default` module as `:shell_default.i`.

```
$ erl
Erlang/OTP 19 [erts-8.1] [source] [64-bit] [smp:4:4] [async-threads:10]
[hipe] [kernel-poll:false]

Eshell V8.1 (abort with ^G)
1> i().


| Pid              | Initial Call             | Heap  | Reds   | Msgs |
|------------------|--------------------------|-------|--------|------|
| Registered       | Current Function         | Stack |        |      |
| <0.0.0>          | otp_ring0:start/2        | 376   | 579    | 0    |
| init             | init:loop/1              | 2     |        |      |
| <0.1.0>          | erts_code_purger:start/0 | 233   | 4      | 0    |
| erts_code_purger | erts_code_purger:loop/0  | 3     |        |      |
| <0.4.0>          | erlang:apply/2           | 987   | 100084 | 0    |
| erl_prim_loader  | erl_prim_loader:loop/3   | 5     |        |      |


```

<0.30.0>	gen_event:init/6	610	226	0
error_logger	gen_event:fetch_msg/5	8		
<0.31.0>	erlang:apply/2	1598	416	0
application_controller	gen_server:loop/6	7		
<0.33.0>	application_master:init/4	233	64	0
	application_master:main_loop/2	6		
<0.34.0>	application_master:start_it/4	233	59	0
	application_master:loop_it/4	5		
<0.35.0>	supervisor:kernel/1	610	1767	0
kernel_sup	gen_server:loop/6	9		
<0.36.0>	erlang:apply/2	6772	73914	0
code_server	code_server:loop/1	3		
<0.38.0>	rpc:init/1	233	21	0
rex	gen_server:loop/6	9		
<0.39.0>	global:init/1	233	44	0
global_name_server	gen_server:loop/6	9		
<0.40.0>	erlang:apply/2	233	21	0
	global:loop_the_locker/1	5		
<0.41.0>	erlang:apply/2	233	3	0
	global:loop_the_registrar/0	2		
<0.42.0>	inet_db:init/1	233	209	0
inet_db	gen_server:loop/6	9		
<0.44.0>	global_group:init/1	233	55	0
global_group	gen_server:loop/6	9		
<0.45.0>	file_server:init/1	233	79	0
file_server_2	gen_server:loop/6	9		
<0.46.0>	supervisor_bridge:standard_error/	233	34	0
standard_error_sup	gen_server:loop/6	9		
<0.47.0>	erlang:apply/2	233	10	0
standard_error	standard_error:server_loop/1	2		
<0.48.0>	supervisor_bridge:user_sup/1	233	54	0
	gen_server:loop/6	9		
<0.49.0>	user_drv:server/2	987	1975	0
user_drv	user_drv:server_loop/6	9		
<0.50.0>	group:server/3	233	40	0
user	group:server_loop/3	4		
<0.51.0>	group:server/3	987	12508	0
	group:server_loop/3	4		
<0.52.0>	erlang:apply/2	4185	9537	0
	shell:shell_rep/4	17		
<0.53.0>	kernel_config:init/1	233	255	0
	gen_server:loop/6	9		
<0.54.0>	supervisor:kernel/1	233	56	0
kernel_safe_sup	gen_server:loop/6	9		
<0.58.0>	erlang:apply/2	2586	18849	0
	c:pinfo/1	50		
Total		23426	220863	0
		222		
ok				

The **i/0** function prints out a list of all processes in the system. Each process gets two lines of information. The first two lines of the printout are the headers telling you what the information means. As you can see you get the Process ID (Pid) and the name of the process if any, as well as information about the code the process is started with and is executing. You also get information about the heap and stack size and the number of reductions and messages in the process. In the rest of this chapter we will learn in detail

what a stack, a heap, a reduction and a message are. For now we can just assume that if there is a large number for the heap size, then the process uses a lot of memory and if there is a large number for the reductions then the process has executed a lot of code.

We can further examine a process with the `i/3` function. Let us take a look at the `code_server` process. We can see in the previous list that the process identifier (pid) of the `code_server` is `<0.36.0>`. By calling `i/3` with the three numbers of the pid we get this information:

```
2> i(0,36,0).
[{registered_name,code_server},
 {current_function,{code_server,loop,1}},
 {initial_call,{erlang,apply,2}},
 {status,waiting},
 {message_queue_len,0},
 {messages,[]},
 {links,[<0.35.0>]},
 {dictionary,[]},
 {trap_exit,true},
 {error_handler,error_handler},
 {priority,normal},
 {group_leader,<0.33.0>},
 {total_heap_size,46422},
 {heap_size,46422},
 {stack_size,3},
 {reductions,93418},
 {garbage_collection,[{max_heap_size,#{error_logger => true,
                                kill => true,
                                size => 0}],
                      {min_bin_vheap_size,46422},
                      {min_heap_size,233},
                      {fullsweep_after,65535},
                      {minor_gcs,0}]},
 {suspending,[]}]
3>
```

We got a lot of information from this call and in the rest of this chapter we will learn in detail what most of these items mean. The first line tells us that the process has been given a name `code_server`. Next we can see which function the process is currently executing or suspended in (`current_function`) and the name of the function that the process started executing in (`initial_call`).

We can also see that the process is suspended waiting for messages (`{status,waiting}`) and that there are no messages in the mailbox (`{message_queue_len,0}`, `{messages,[]}`). We will look closer at how message passing works later in this chapter.

The fields `priority`, `suspending`, `reductions`, `links`, `trap_exit`, `error_handler`, and `group_leader` control the process execution, error handling, and IO. We will look into this a bit more when we introduce the *Observer*.

The last few fields (`dictionary`, `total_heap_size`, `heap_size`, `stack_size`, and `garbage_collection`) give us information about the process memory usage. We will look at the process memory areas in detail in chapter [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#).

Another, even more intrusive way of getting information about processes is to use the process information given by the `BREAK` menu: `ctrl+c p [enter]`. Note that while you are in the `BREAK` state the whole node freezes.

### 3.1.2. Programmatic Process Probing

The shell functions just print the information about the process but you can actually get this information as data, so you can write

your own tools for inspecting processes. You can get a list of all processes with `erlang:processes/0`, and more information about a process with `erlang:process_info/1`. We can also use the function `whereis/1` to get a pid from a name:

```

1> Ps = erlang:processes().
[<0.0.0>,<0.1.0>,<0.4.0>,<0.30.0>,<0.31.0>,<0.33.0>,
 <0.34.0>,<0.35.0>,<0.36.0>,<0.38.0>,<0.39.0>,<0.40.0>,
 <0.41.0>,<0.42.0>,<0.44.0>,<0.45.0>,<0.46.0>,<0.47.0>,
 <0.48.0>,<0.49.0>,<0.50.0>,<0.51.0>,<0.52.0>,<0.53.0>,
 <0.54.0>,<0.60.0>]
2> CodeServerPid = whereis(code_server).
<0.36.0>
3> erlang:process_info(CodeServerPid).
[{registered_name,code_server},
 {current_function,{code_server,loop,1}},
 {initial_call,{erlang,apply,2}},
 {status,waiting},
 {message_queue_len,0},
 {messages,[]},
 {links,[<0.35.0>]},
 {dictionary,[]},
 {trap_exit,true},
 {error_handler,error_handler},
 {priority,normal},
 {group_leader,<0.33.0>},
 {total_heap_size,24503},
 {heap_size,6772},
 {stack_size,3},
 {reductions,74260},
 {garbage_collection,[{max_heap_size,#{error_logger => true,
                                         kill => true,
                                         size => 0}},,
                      {min_bin_vheap_size,46422},
                      {min_heap_size,233},
                      {fullsweep_after,65535},
                      {minor_gcs,33}]}],
{suspending,[]}]

```

By getting process information as data we can write code to analyze or sort the data as we please. If we grab all processes in the system (with `erlang:processes/0`) and then get information about the heap size of each process (with `erlang:process_info(P, total_heap_size)`) we can then construct a list with pid and heap size and sort it on heap size:

```

1> lists:reverse(lists:keysort(2,[{P,element(2,
    erlang:process_info(P,total_heap_size))}| P <- erlang:processes()]}).
[{{0.36, 24503},
 {0.52, 21916},
 {0.4, 12556},
 {0.58, 4184},
 {0.51, 4184},
 {0.31, 3196},
 {0.49, 2586},
 {0.35, 1597},
 {0.3, 986},
 {0.0, 752},
 {0.33, 609},
 {0.54, 233},
 {0.53, 233},
 {0.5, 233},
 {0.48, 233},
 {0.47, 233},
 {0.46, 233},
 {0.45, 233},
 {0.44, 233},
 {0.42, 233},
 {0.41, 233},
 {0.4, 233},
 {0.39, 233},
 {0.38, 233},
 {0.34, 233},
 {0.1, 233}]
2>

```

You might notice that many processes have a heap size of 233, that is because it is the default starting heap size of a process.

See the documentation of the module `erlang` for a full description of the information available with `process_info` [[http://erlang.org/doc/man/erlang.html#process\\_info-1](http://erlang.org/doc/man/erlang.html#process_info-1)]. Notice how the `process_info/1` function only returns a subset of all the information available for the process and how the `process_info/2` function can be used to fetch extra information. As an example, to extract the `backtrace` for the `code_server` process above, we could run:

```

3> process_info(whereis(code_server), backtrace).
{backtrace,<<"Program counter: 0x00000000161de900 (code_server:loop/1 + 152)\nCP: 0x0000000000000000\n(invalid)\narity = 0\n\n0"...
>>}

```

See the three dots at the end of the binary above? That means that the output has been truncated. A useful trick to see the whole value is to wrap the above function call using the `rp/1` function:

```

4> rp(process_info(whereis(code_server), backtrace)).

```

An alternative is to use the `io:put_chars/1` function, as follows:

```
5> {backtrace, Backtrace} = process_info(whereis(code_server), backtrace).
{backtrace,<<"Program counter: 0x00000000161de900 (code_server:loop/1 + 152)\nCP: 0x0000000000000000
(invalid)\narity = 0\n\n0"...
6> io:put_chars(Backtrace).
```

Due to its verbosity, the output for commands 4> and 6> has not been included here, but feel free to try the above commands in your Erlang shell.

### 3.1.3. Using the Observer to Inspect Processes

A third way of examining processes is with the *Observer* [[http://erlang.org/doc/apps/observer/observer\\_ug.html](http://erlang.org/doc/apps/observer/observer_ug.html)]. The Observer is an extensive graphical interface for inspecting the Erlang RunTime System. We will use the Observer throughout this book to examine different aspects of the system.

The Observer can either be started from the OS shell and attach itself to an node or directly from an Elixir or Erlang shell. For now we will just start the Observer from the Elixir shell with `:observer.start` or from the Erlang shell with:

```
7> observer:start().
```

When the Observer is started it will show you a system overview, see the following screen shot:

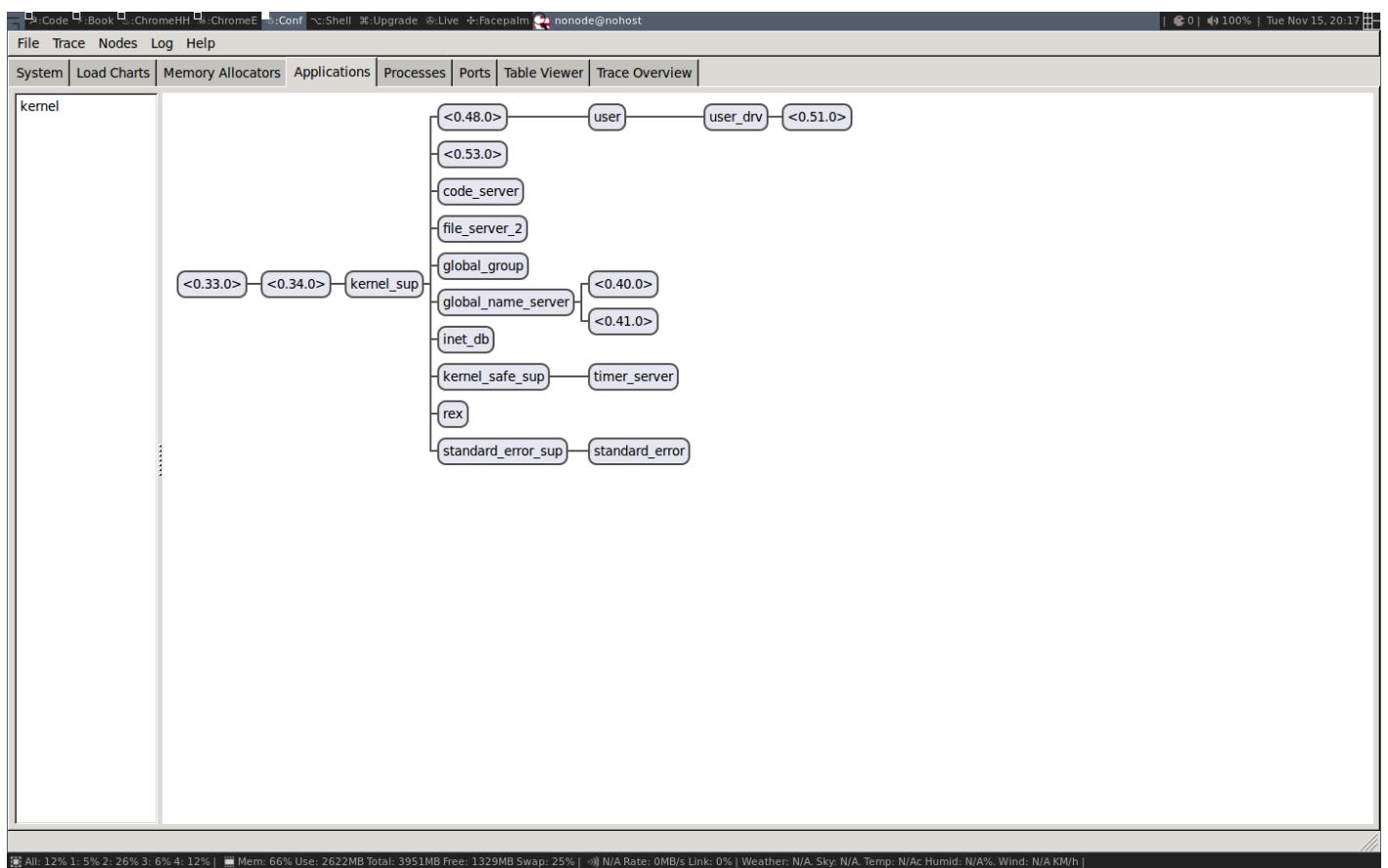
System Version:	19
ERTS Version:	8.1
Compiled for:	x86_64-unknown-linux-gnu
Emulator Wordsize:	8
Process Wordsize:	8
SMP Support:	true
Thread Support:	true
Sync thread pool size:	10

Total:	18 MB
Processes:	5342 kB
Atoms:	331 kB
Binaries:	296 kB
Code:	6628 kB
ETS:	658 kB

Logical CPU's:	4
Online Logical CPU's:	4
Available Logical CPU's:	4
Schedulers:	4
Online schedulers:	4
Available schedulers:	4

Up time:	45 Mins
Max Processes:	262144
Processes:	44
Run Queue:	0
IO Input:	5082 kB
IO Output:	55 kB

We will go over some of this information in detail later in this and the next chapter. For now we will just use the Observer to look at the running processes. First we take a look at the `Applications` tab which shows the supervision tree of the running system:



Here we get a graphical view of how the processes are linked. This is a very nice way to get an overview of how a system is structured. You also get a nice feeling of processes as isolated entities floating in space connected to each other through links.

To actually get some useful information about the processes we switch to the **Processes** tab:

## The Erlang Runtime System Part I: Understanding ERTS

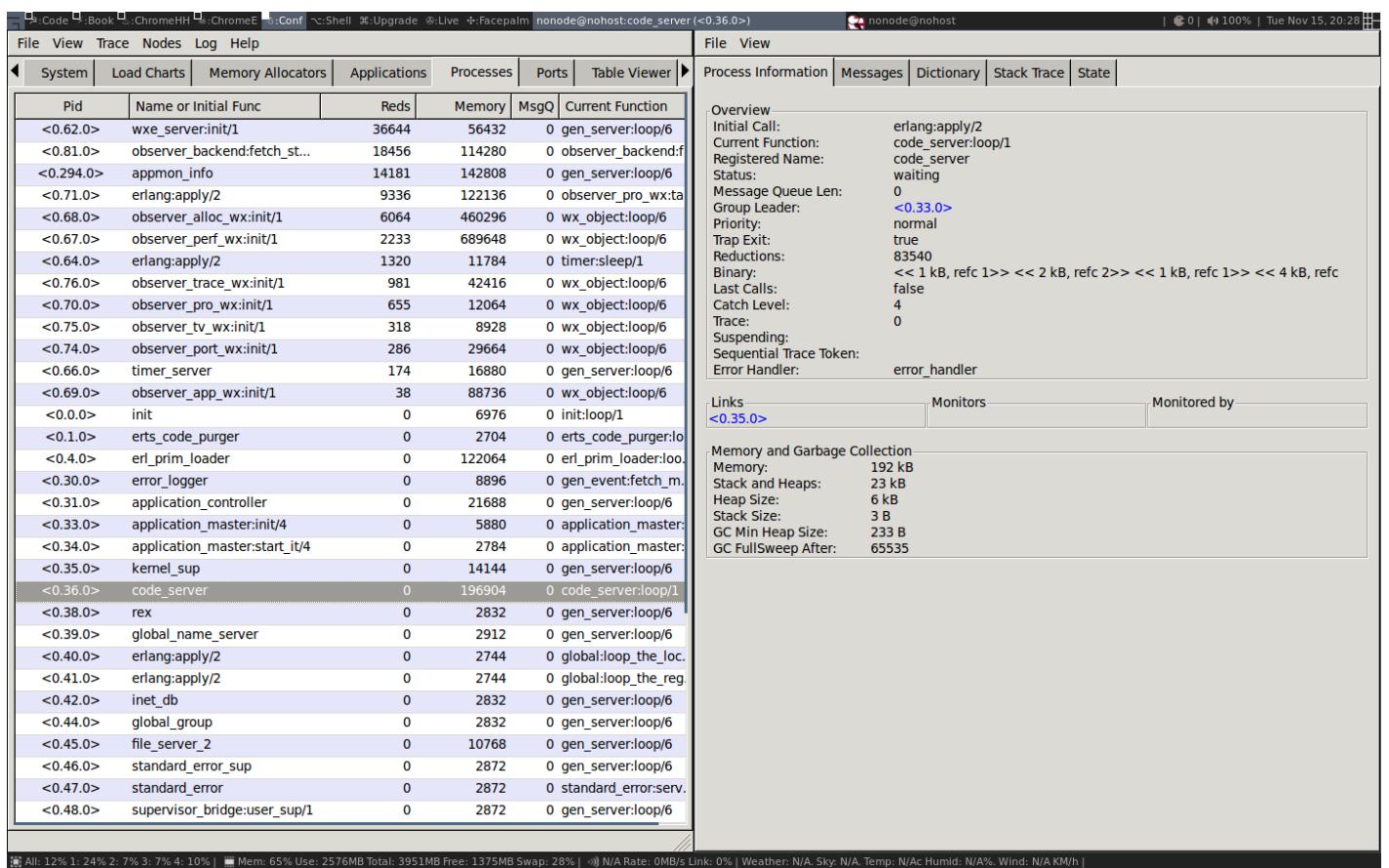
The screenshot shows a web-based interface for monitoring Erlang processes. At the top, there's a navigation bar with links like File, View, Trace, Nodes, Log, Help, System, Load Charts, Memory Allocators, Applications, Processes, Ports, Table Viewer, and Trace Overview. Below the navigation bar is a table with the following columns: Pid, Name or Initial Func, Reds, Memory, MsgQ, and Current Function. The table lists numerous processes, each with its name, reduction count, memory usage, message queue count, and the current function it's executing. One row, specifically the entry for 'erl\_prim\_loader', is highlighted with a dark grey background.

Pid	Name or Initial Func	Reds	Memory	MsgQ	Current Function
<0.81.0>	observer_backend:fetch_st...	18456	109256	0	observer_backend:fetch_stats_loop/2
<0.62.0>	wxe_server:init/1	14093	17984	0	gen_server:loop/6
<0.294.0>	appmon_info	13360	142808	0	gen_server:loop/6
<0.68.0>	observer_alloc_wx:init/1	5766	460584	0	wx_object:loop/6
<0.71.0>	erlang:apply/2	4001	142832	0	observer_pro_wx:table_holder/1
<0.67.0>	observer_perf_wx:init/1	2193	176408	0	wx_object:loop/6
<0.64.0>	erlang:apply/2	1320	11784	0	timer:sleep/1
<0.66.0>	timer_server	186	24744	0	gen_server:loop/6
<0.70.0>	observer_pro_wx:init/1	72	12024	0	wx_object:loop/6
<0.0.0>	init	0	6976	0	init:loop/1
<0.1.0>	erts_code_purger	0	2704	0	erts_code_purger:loop/0
<0.4.0>	erl_prim_loader	0	142760	0	erl_prim_loader:loop/3
<0.30.0>	error_logger	0	8896	0	gen_event:fetch_msg/5
<0.31.0>	application_controller	0	26576	0	gen_server:loop/6
<0.33.0>	application_master:init/4	0	5880	0	application_master:main_loop/2
<0.34.0>	application_master:start_it/4	0	2784	0	application_master:loop_it/4
<0.35.0>	kernel_sup	0	14144	0	gen_server:loop/6
<0.36.0>	code_server	0	196904	0	code_server:loop/1
<0.38.0>	rex	0	2832	0	gen_server:loop/6
<0.39.0>	global_name_server	0	2912	0	gen_server:loop/6
<0.40.0>	erlang:apply/2	0	2744	0	global:loop_the_locker/1
<0.41.0>	erlang:apply/2	0	2744	0	global:loop_the_registrar/0
<0.42.0>	inet_db	0	2832	0	gen_server:loop/6
<0.44.0>	global_group	0	2832	0	gen_server:loop/6
<0.45.0>	file_server_2	0	10768	0	gen_server:loop/6
<0.46.0>	standard_error_sup	0	2872	0	gen_server:loop/6
<0.47.0>	standard_error	0	2872	0	standard_error:server_loop/1
<0.48.0>	supervisor_bridge:user_sup/1	0	2872	0	gen_server:loop/6
<0.49.0>	user_drv	0	68008	0	user_drv:server_loop/6
<0.50.0>	user	0	2912	0	group:server_loop/3
<0.51.0>	group:server/3	0	34552	0	group:more_data/5
<0.52.0>	erlang:apply/2	0	176376	0	shell:get_command1/5

All: 11% 1: 13% 2: 15% 3: 11% 4: 5% | Mem: 66% Use: 2620MB Total: 3951MB Free: 1331MB Swap: 25% | N/A Rate: 0MB/s Link: 0% | Weather: N/A Sky: N/A Temp: N/A Humid: N/A% Wind: N/A KM/h

In this view we get basically the same information as with `i/0` in the shell. We see the pid, the registered name, number of reductions, memory usage and number of messages and the current function.

We can also look into a process by double clicking on its row, for example on the code server, to get the kind of information you can get with `process_info/2`:



All: 12% 1: 24% 2: 7% 3: 7% 4: 10% Mem: 65% Use: 2576MB Total: 3951MB Swap: 28% N/A Rate: 0MB/s Link: 0% Weather: N/A Sky: N/A Temp: N/A Humid: N/A% Wind: N/A KM/h

We will not go through what all this information means right now, but if you keep on reading all will eventually be revealed.

## Enabling the Observer

If you are building your application with erlang.mk or rebar and you want to include the Observer application in your build you might need to add the applications `runtime_tools`, `wx`, and `observer` to your list of applications in yourapp.app.src.

Now that we have a basic understanding of what a process is and some tools to find and inspect processes in a system we are ready to dive deeper to learn how a process is implemented.

## 3.2. Processes Are Just Memory

A process is basically four blocks of memory: a *stack*, a *heap*, a *message area*, and the *Process Control Block (the PCB)*.

The stack is used for keeping track of program execution by storing return addresses, for passing arguments to functions, and for keeping local variables. Larger structures, such as lists and tuples are stored on the heap.

The *message area*, also called *the mailbox*, is used to store messages sent to the process from other processes. The process control block is used to keep track of the state of the process.

See the following figure for an illustration of a process as memory:



*Figure 7. Erlang Process Memory : Basic*

This picture of a process is very much simplified, and we will go through a number of iterations of more refined versions to get to a more accurate picture.

The stack, the heap, and the mailbox are all dynamically allocated and can grow and shrink as needed. We will see exactly how this works in later chapters. The PCB on the other hand is statically allocated and contains a number of fields that controls the process.

We can actually inspect some of these memory areas by using *HiPE’s Built In Functions* (HiPE BIFs) for introspection. With these BIFs we can print out the memory content of stacks, heaps, and the PCB. The raw data is printed and in most cases a human readable version is pretty printed alongside the data. To really understand everything that we see when we inspect the memory we will need to know more about the Erlang tagging scheme (which we will go through in [The Erlang Type System and Tags](#)) and about the execution model and error handling which we will go through in [The Erlang Virtual Machine: BEAM](#), but using these tools will give us a nice view of how a process really is just memory.

### HiPE’s Built In Functions (HiPE BIFs)

The HiPE BIFs are not an official part of Erlang/OTP. They are not supported by the OTP team. They might be removed or changed at any time, so don’t base your mission critical services on them.

These BIFs examine the internals of ERTS in a way that might not be safe. The BIFs for introspection often just print to standard out and you might be surprised where that output ends up.

These BIFs can lock up a scheduler thread for a long time without using any reductions (we will look at what that means in the next chapter). Printing the heap of a very large process for example can take a long time.

These BIFs are only meant to be used for debugging and you use them at your own risk. You should probably not run them on a live system.

Many of the HiPE BIFs were written by the author in the mid nineties (before 64 bit Erlang existed) and the printouts on a 64 bit machine might be a bit off. There are new versions of these BIFs that do a better job, hopefully they will be included in ERTS at the time of the printing of this book. Otherwise you can build your own version with the patch provided in the code section and the instructions in [\[CH-BuildingERTS\]](#).

We can see the context of the stack of a process with `hipec_bifs:show_estack/1`:

```

1> hipe_bifs:show_estack(self()).
   BEAM  STACK
   Address |   Contents |
-----|-----
 0x00007f9cc3238310 | 0x00007f9cc2ea6fe8 | BEAM ACTIVATION RECORD
 0x00007f9cc3238318 | 0xfffffffffffffffffb | BEAM PC shell:exprs/7 + 0x4e
 0x00007f9cc3238320 | 0x000000000000644b | []
-----|-----
 0x00007f9cc3238328 | 0x00007f9cc2ea6708 | BEAM ACTIVATION RECORD
 0x00007f9cc3238330 | 0xfffffffffffffffffb | BEAM PC shell:eval_exprs/7 + 0xf
 0x00007f9cc3238338 | 0xfffffffffffffffffb | []
 0x00007f9cc3238340 | 0x000000000004f3cb | cmd
 0x00007f9cc3238348 | 0xfffffffffffffffffb | []
 0x00007f9cc3238350 | 0x00007f9cc3237102 | {value,#Fun<shell.5.104321512>}
 0x00007f9cc3238358 | 0x00007f9cc323711a | {eval,#Fun<shell.21.104321512>}
 0x00007f9cc3238360 | 0x00000000000200ff | 8207
 0x00007f9cc3238368 | 0xfffffffffffffffffb | []
 0x00007f9cc3238370 | 0xfffffffffffffffffb | []
 0x00007f9cc3238378 | 0xfffffffffffffffffb | []
-----|-----
 0x00007f9cc3238380 | 0x00007f9cc2ea6300 | BEAM ACTIVATION RECORD
 0x00007f9cc3238388 | 0xfffffffffffffffffb | BEAM PC shell:eval_loop/3 + 0x47
 0x00007f9cc3238390 | 0xfffffffffffffffffb | []
 0x00007f9cc3238398 | 0xfffffffffffffffffb | []
 0x00007f9cc32383a0 | 0xfffffffffffffffffb | []
 0x00007f9cc32383a8 | 0x00001a00000343 | <0.52.0>
-----|-----
 0x00007f9cc32383b0 | 0x000000000005a9b | BEAM CATCH FRAME
                           | CATCH 0x00007f9cc2ea67d8
                           | (BEAM shell:exprs/7 + 0x29)
*****|*****
-----|-----
 0x00007f9cc32383b8 | 0x000000000093aeb8 | BEAM ACTIVATION RECORD
 0x00007f9cc32383c0 | 0x000000000000200ff | BEAM PC normal-process-exit
 0x00007f9cc32383c8 | 0x00001a00000343 | 8207
-----|-----

```

true  
2>

We will look closer at the values on the stack and the heap in [The Erlang Type System and Tags](#). The content of the heap is printed by `hipe_bifs:show_heap/1`. Since we do not want to list a large heap here we'll just spawn a new process that does nothing and show that heap:

```

2> hipe_bifs:show_heap(spawn(fun () -> ok end)).
From: 0x00007f7f33ec9588 to 0x00007f7f33ec9848
| H E A P |
| Address | Contents |
+-----+
| 0x00007f7f33ec9588 | 0x00007f7f33ec959a | #Fun<erl_eval.20.52032458>
| 0x00007f7f33ec9590 | 0x00007f7f33ec9839 | []
| 0x00007f7f33ec9598 | 0x00000000000000154 | Thing Arity(5) Tag(20)
| 0x00007f7f33ec95a0 | 0x00007f7f3d3833d0 | THING
| 0x00007f7f33ec95a8 | 0x0000000000000000 | THING
| 0x00007f7f33ec95b0 | 0x000000000000600324 | THING
| 0x00007f7f33ec95b8 | 0x0000000000000000 | THING
| 0x00007f7f33ec95c0 | 0x0000000000000001 | THING
| 0x00007f7f33ec95c8 | 0x000001d0000003a3 | <0.58.0>
| 0x00007f7f33ec95d0 | 0x00007f7f33ec95da | {[],{eval...
| 0x00007f7f33ec95d8 | 0x00000000000000100 | Arity(4)
| 0x00007f7f33ec95e0 | 0xfffffffffffffffffb | []
| 0x00007f7f33ec95e8 | 0x00007f7f33ec9602 | {eval,#Fun<shell.21.104321512>}
| 0x00007f7f33ec95f0 | 0x00007f7f33ec961a | {value,#Fun<shell.5.104321512>}...
| 0x00007f7f33ec95f8 | 0x00007f7f33ec9631 | [{clause...
...
| 0x00007f7f33ec97d0 | 0x00007f7f33ec97fa | #Fun<shell.5.104321512>
| 0x00007f7f33ec97d8 | 0x000000000000000c0 | Arity(3)
| 0x00007f7f33ec97e0 | 0x00000000000000e4b | atom
| 0x00007f7f33ec97e8 | 0x0000000000000001f | 1
| 0x00007f7f33ec97f0 | 0x0000000000006d0b | ok
| 0x00007f7f33ec97f8 | 0x00000000000000154 | Thing Arity(5) Tag(20)
| 0x00007f7f33ec9800 | 0x00007f7f33bde0c8 | THING
| 0x00007f7f33ec9808 | 0x00007f7f33ec9780 | THING
| 0x00007f7f33ec9810 | 0x000000000060030c | THING
| 0x00007f7f33ec9818 | 0x00000000000000002 | THING
| 0x00007f7f33ec9820 | 0x00000000000000001 | THING
| 0x00007f7f33ec9828 | 0x000001d0000003a3 | <0.58.0>
| 0x00007f7f33ec9830 | 0x000001a000000343 | <0.52.0>
| 0x00007f7f33ec9838 | 0xfffffffffffffb | []
| 0x00007f7f33ec9840 | 0xfffffffffffffb | []
+-----+
true
3>

```

We can also print the content of some of the fields in the PCB with `hipe\_bifs:show\_pcb/1`:

```

3> hipe_bifs:show_pcb(self()).
P: 0x00007f7f3cbc0400
-----+
Offset| Name      | Value      | *Value
      | id        | 0x000001d0000003a3 |
      | htop      | 0x00007f7f33f15298 |
      | hend      | 0x00007f7f33f16540 |
      | heap      | 0x00007f7f33f11470 |
-----+

```

104	heap_sz	0x000000000000000a1a
80	stop	0x00007f7f33f16480
592	gen_gcs	0x00000000000000012
594	max_gen_gcs	0x000000000000ffff
552	high_water	0x00007f7f33f11c50
560	old_hend	0x00007f7f33e90648
568	old_htop	0x00007f7f33e8f8e8
576	old_head	0x00007f7f33e8e770
112	min_heap...	0x00000000000000e9
328	rcount	0x0000000000000000
336	reds	0x0000000000002270
16	tracer	0xfffffffffffffffffb
24	trace_fla...	0x0000000000000000
344	group_lea...	0x0000019800000333
352	flags	0x0000000000002000
360	fvalue	0xfffffffffffffffffb
368	freason	0x0000000000000000
320	fcalls	0x00000000000005a2
384	next	0x0000000000000000
48	reg	0x0000000000000000
56	nlinks	0x00007f7f3cbc0750
616	mbuf	0x0000000000000000
640	mbuf_sz	0x0000000000000000
464	dictionary	0x0000000000000000
472	seq..clock	0x0000000000000000
480	seq..astcnt	0x0000000000000000
488	seq..token	0xfffffffffffffb
496	intial[0]	0x000000000000320b
504	intial[1]	0x000000000000c8b
512	intial[2]	0x000000000000002
520	current	0x00007f7f3be87c20
296	cp	0x00007f7f3d3a5100
304	i	0x00007f7f3be87c38
312	catches	0x0000000000000001
224	arity	0x0000000000000000
232	arg_reg	0x00007f7f3cbc04f8
240	max_arg_reg	0x0000000000000006
248	def..reg[0]	0x000000000000320b
256	def..reg[1]	0x000000000000c8b
264	def..reg[2]	0x00007f7f33ec9589
272	def..reg[3]	0x0000000000000000
280	def..reg[4]	0x0000000000000000
288	def..reg[5]	0x00000000000007d0
136	nsp	0x0000000000000000
144	nstack	0x0000000000000000
152	nstend	0x0000000000000000
160	ncallee	0x0000000000000000
56	ncsp	0x0000000000000000
64	narity	0x0000000000000000

true

4&gt;

Now armed with these inspection tools we are ready to look at what these fields in the PCB mean.

### 3.3. The PCB

The Process Control Block contains all the fields that control the behaviour and current state of a process. In this section and the rest of the chapter we will go through the most important fields. We will leave out some fields that have to do with execution and tracing from this chapter, instead we will cover those in [The Erlang Virtual Machine: BEAM](#).

If you want to dig even deeper than we will go in this chapter you can look at the C source code. The PCB is implemented as a C struct called `process` in the file `erl_process.h` [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/erl\\_process.h](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/erl_process.h)].

The field `id` contains the process ID (or PID).

0   id	0x000001d0000003a3	
--------	--------------------	--

The process ID is an Erlang term and hence tagged (See [The Erlang Type System and Tags](#)). This means that the 4 least significant bits are a tag (0011). In the code section there is a module for inspecting Erlang terms (see `show.erl`) which we will cover in the chapter on types. We can use it now to examine the type of a tagged word though.

```
4> show:tag_to_type(16#0000001d0000003a3).
pid
5>
```

The fields `hend` and `stop` are pointers to the top of the heap and the stack, that is, they are pointing to the next free slots on the heap or stack. The fields `heap` (start) and `hend` points to the start and the stop of the whole heap, and `heap_sz` gives the size of the heap in words. That is `hend - heap = heap_sz * 8` on a 64 bit machine and `hend - heap = heap_sz * 4` on a 32 bit machine.

The field `min_heap_size` is the size, in words, that the heap starts with and which it will not shrink smaller than, the default value is 233.

We can now refine the picture of the process heap with the fields from the PCB that controls the shape of the heap:

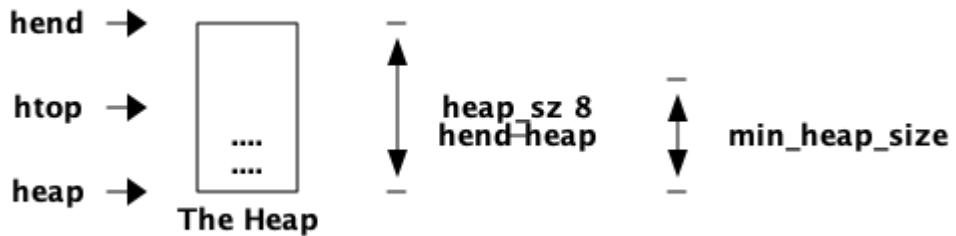


Figure 8. Erlang Process Heap

But wait, how come we have a heap start and a heap end, but no start and stop for the stack? That is because the BEAM uses a trick to save space and pointers by allocating the heap and the stack together. It is time for our first revision of our process as memory picture. The heap and the stack are actually just one memory area:



Figure 9. Erlang Process Memory : Heap + Stack

The stack grows towards lower memory addresses and the heap towards higher memory, so we can also refine the picture of the heap by adding the stack top pointer to the picture:

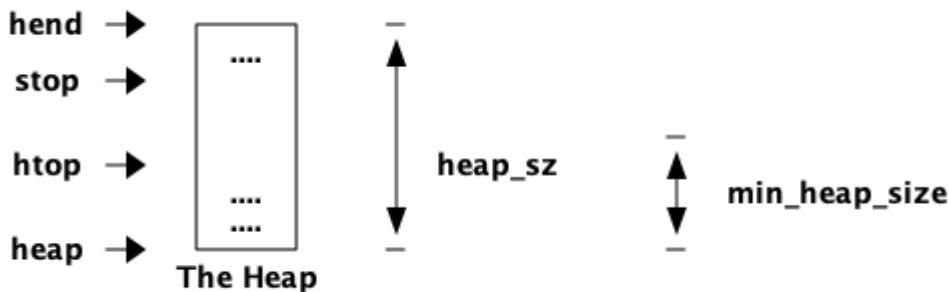


Figure 10. Erlang Process Heap and Stack

If the pointers **htop** and **stop** were to meet, the process would run out of free memory and would have to do a garbage collection to free up memory.

### 3.4. The Garbage Collector (GC)

The heap memory management schema is to use a per process copying generational garbage collector. When there is no more space on the heap (or the stack, since they share the allocated memory block), the garbage collector kicks in to free up memory.

The GC allocates a new memory area called the *to space*. Then it goes through the stack to find all live roots and follows each root and copies the data on the heap to the new heap. Finally it also copies the stack to the new heap and frees up the old memory area.

The GC is controlled by these fields in the PCB:

```
Eterm *high_water;
Eterm *old_hend; /* Heap pointers for generational GC. */
Eterm *old_htop;
Eterm *old_heap;
Uint max_heap_size; /* Maximum size of heap (in words). */
Uint16 gen_gcs; /* Number of (minor) generational GCs. */
Uint16 max_gen_gcs; /* Max minor gen GCs before fullsweep. */
```

Since the garbage collector is generational it will use a heuristic to just look at new data most of the time. That is, in what is called a *minor collection*, the GC only looks at the top part of the stack and moves new data to the new heap. Old data, that is data allocated below the **high\_water** mark (see the figure below) on the heap, is moved to a special area called the old heap.

Most of the time, then, there is another heap area for each process: the old heap, handled by the fields **old\_heap**, **old\_htop** and **old\_hend** in the PCB. This almost brings us back to our original picture of a process as four memory areas:



Figure 11. Erlang Process Memory : GC

When a process starts there is no old heap, but as soon as young data has matured to old data and there is a garbage collection, the old heap is allocated. The old heap is garbage collected when there is a *major collection*, also called a *full sweep*. See [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#) for more details of how garbage collection works. In that chapter we will also look at how to track down and fix memory related problems.

## 3.5. Mailboxes and Message Passing

Process communication is done through message passing. A process send is implemented so that a sending process copies the message from its own heap to the mailbox of the receiving process.

In the early days of Erlang concurrency was implemented through multitasking in the scheduler. We will talk more about concurrency in the section about the scheduler later in this chapter, for now it is worth noting that in the first version of Erlang there was no parallelism and there could only be one process running at the time. In that version the sending process could write data directly on the receiving process' heap.

### 3.5.1. Sending Messages in Parallel

When multicore systems were introduced and the Erlang implementation was extended with several schedulers running processes in parallel it was no longer safe to write directly on another process' heap without taking the **main lock** of the receiver. At this time the concept of **m-bufs** was introduced (also called **heap fragments**). An **m-buf** is a memory area outside of a process heap where other processes can safely write data. If a sending process can not get the lock it would write to the **m-buf** instead. When all data of a message has been copied to the **m-buf** the message is linked to the process through the mailbox. The linking (*LINK\_MESSAGE* in `erl_message.h` [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/erl\\_message.h](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/erl_message.h)]) appends the message to the receiver's message queue.

The garbage collector would then copy the messages onto the process' heap. To reduce the pressure on the GC the mailbox is divided into two lists, one containing seen messages and one containing new messages. The GC does not have to look at the new messages since we know they will survive (they are still in the mailbox) and that way we can avoid some copying.

## 3.6. Lock Free Message Passing

In Erlang 19 a new per process setting was introduced, `message_queue_data`, which can take the values `on_heap` or `off_heap`. When set to `on_heap` the sending process will first try to take the **main lock** of the receiver and if it succeeds the message will be copied directly onto the receiver's heap. This can only be done if the receiver is suspended and if no other process has grabbed the lock to send to the same process. If the sender can not obtain the lock it will allocate a heap fragment and copy the message there instead.

If the flag is set to `off_heap` the sender will not try to get the lock and instead write directly to a heap fragment. This will reduce lock contention but allocating a heap fragment is more expensive than writing directly to the already allocated process heap and it can lead to larger memory usage. There might be a large empty heap allocated and still new messages are written to new fragments.

With `on_heap` allocation all the messages, both directly allocated on the heap and messages in heap fragments, will be copied by the GC. If the message queue is large and many messages are not handled and therefore still are live, they will be promoted to the old heap and the size of the process heap will increase, leading to higher memory usage.

All messages are added to a linked list (the mailbox) when the message has been copied to the receiving process. If the message is copied to the heap of the receiving process the message is linked in to the **internal message queue** (or **seen** messages) and examined by the GC. In the *off\_heap* allocation scheme new messages are placed in the "external" **message in queue** and ignored by the GC.

### 3.6.1. Memory Areas for Messages

We can now revise our picture of the process as four memory areas once more. Now the process is made up of five memory areas (two mailboxes) and a varying number of heap fragments (**m-bufs**):

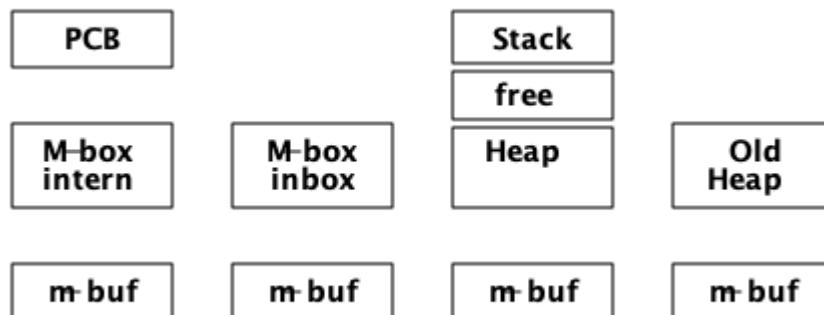


Figure 12. Erlang Process Memory : Messages

Each mailbox consists of a length and two pointers, stored in the fields `msg.len`, `msg.first`, `msg.last` for the internal queue and `msg_inq.len`, `msg_inq.first`, and `msg_inq.last` for the external in queue. There is also a pointer to the next message to look at (`msg.save`) to implement selective receive.

### 3.6.2. Inspecting Message Handling

Let us use our introspection tools to see how this works in more detail. We start by setting up a process with a message in the mailbox and then take a look at the PCB.

```

4> P = spawn(fun() -> receive stop -> ok end end).
<0.63.0>
5> P ! start.
start
6> hipe_bifs:show_pcb(P).

```

```

...
408 | msg.first      | 0x00007fd40962d880 |
416 | msg.last       | 0x00007fd40962d880 |
424 | msg.save        | 0x00007fd40962d880 |
432 | msg.len         | 0x0000000000000001 |
696 | msg_inq.first   | 0x0000000000000000 |
704 | msg_inq.last    | 0x00007fd40a306238 |
712 | msg_inq.len     | 0x0000000000000000 |
616 | mbuf            | 0x0000000000000000 |
640 | mbuf_sz         | 0x0000000000000000 |
...

```

From this we can see that there is one message in the message queue and the `first`, `last` and `save` pointers all point to this

message.

As mentioned we can force the message to end up in the in queue by setting the flag `message_queue_data`. We can try this with the following program:

```
-module(msg).

-export([send_on_heap/0
        ,send_off_heap/0]).

send_on_heap() -> send(on_heap).
send_off_heap() -> send(off_heap).

send(How) ->
    %% Spawn a function that loops for a while
    P2 = spawn(fun () -> receiver(How) end),
    %% spawn a sending process
    P1 = spawn(fun () -> sender(P2) end),
    P1.

sender(P2) ->
    %% Send a message that ends up on the heap
    %% {_,S} = erlang:process_info(P2, heap_size),
    M = loop(0),
    P2 ! self(),
    receive ready -> ok end,
    P2 ! M,
    %% Print the PCB of P2
    hipe_bifs:show_pcb(P2),
    ok.

receiver(How) ->
    erlang:process_flag(message_queue_data, How),
    receive P -> P ! ready end,
    %% loop(100000),
    receive x -> ok end,
    P.

loop(0) -> [done];
loop(N) -> [loop(N-1)].
```

With this program we can try sending a message on heap and off heap and look at the PCB after each send. With on heap we get the same result as when just sending a message before:

```
5> msg:send_on_heap().
```

...

408	msg.first	0x00007fd4096283c0	
416	msg.last	0x00007fd4096283c0	
424	msg.save	0x00007fd40a3c1048	
432	msg.len	0x0000000000000001	
696	msg_inq.first	0x0000000000000000	
704	msg_inq.last	0x00007fd40a3c1168	
712	msg_inq.len	0x0000000000000000	
616	mbuf	0x0000000000000000	
640	mbuf_sz	0x0000000000000000	

...

If we try sending to a process with the flag set to *off\_heap* the message ends up in the in queue instead:

```
6> msg:send_off_heap().
```

...

408	msg.first	0x0000000000000000	
416	msg.last	0x00007fd40a3c0618	
424	msg.save	0x00007fd40a3c0618	
432	msg.len	0x0000000000000000	
696	msg_inq.first	0x00007fd3b19f1830	
704	msg_inq.last	0x00007fd3b19f1830	
712	msg_inq.len	0x0000000000000001	
616	mbuf	0x0000000000000000	
640	mbuf_sz	0x0000000000000000	

...

### 3.6.3. The Process of Sending a Message to a Process

We will ignore the distribution case for now, that is we will not consider messages sent between Erlang nodes. Imagine two processes **P1** and **P2**. Process **P1** wants to send a message (*Msg*) to process **P2**, as illustrated by this figure:

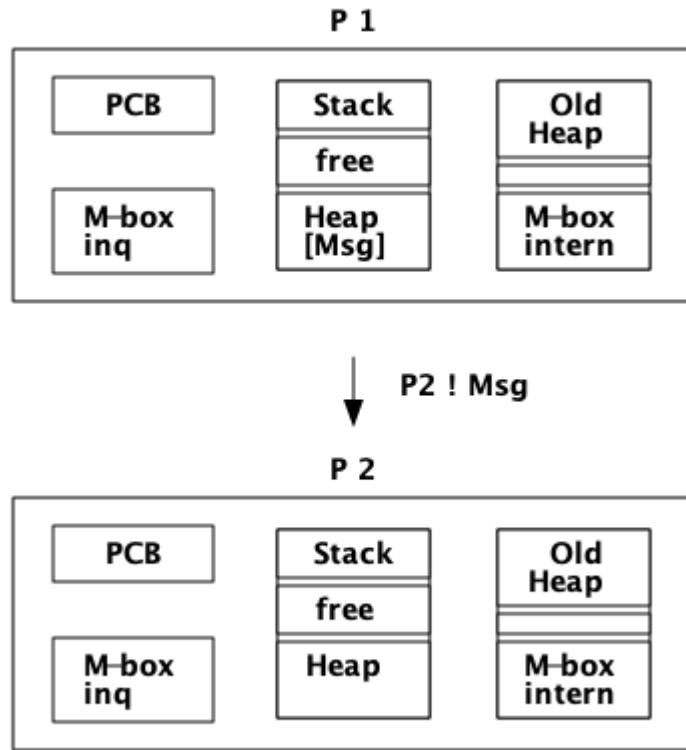


Figure 13. Erlang Message Passing Step 1

Process P1 will then take the following steps:

- Calculate the size of *Msg*.
- Allocate space for the message (on or off P2's heap as described before).
- Copy *Msg* from P1's heap to the allocated space.
- Allocate and fill in an *ErlMessage* struct wrapping up the message.
- Link in the *ErlMessage* either in the *ErlMsgQueue* or in the *ErlMsgInQueue*.

If process P2 is suspended and no other process is trying to send a message to P2 and there is space on the heap and the allocation strategy is *on\_heap* the message will directly end up on the heap:

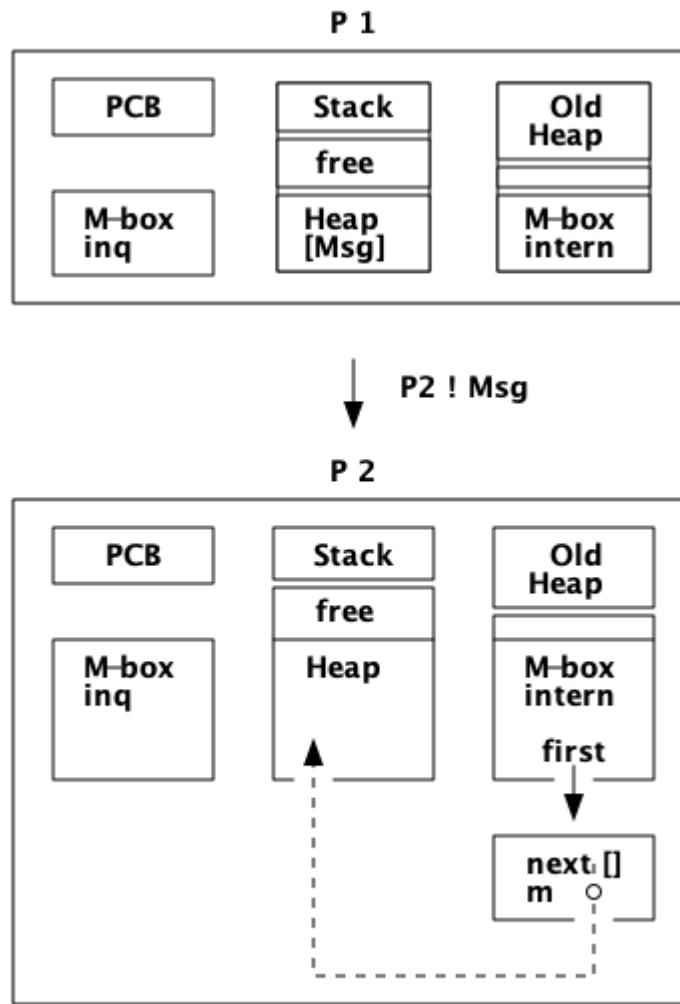


Figure 14. Erlang Message Passing Step 2

If P1 can not get the `main lock` of P2 or there is not enough space on P2's heap and the allocation strategy is `on_heap` the message will end up in an `m-buf` but linked from the internal mailbox:

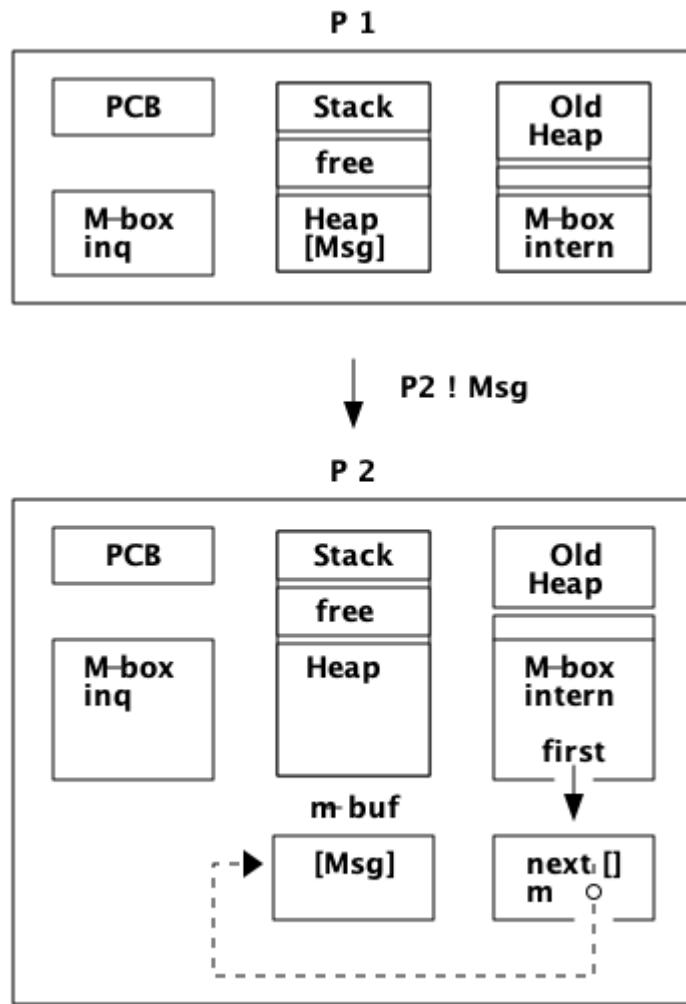


Figure 15. Erlang Message Passing Step 3

After a GC the message will be moved into the heap.

If the allocation strategy is *off\_heap* the message will end up in an **m-buf** and linked from the external mailbox:

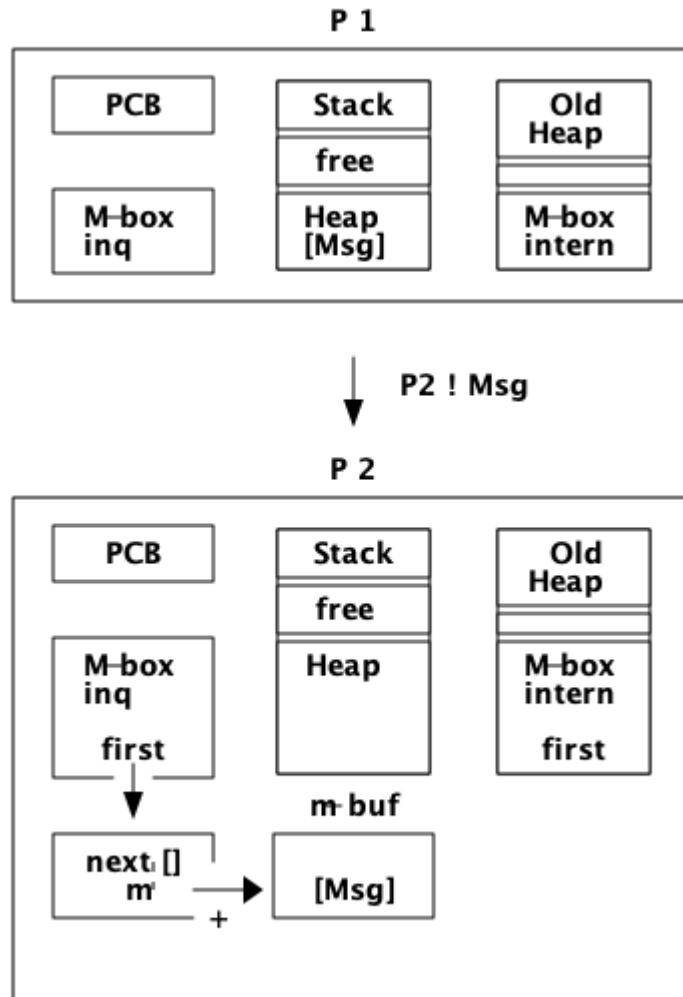


Figure 16. Erlang Message Passing Step 4

After a GC the message will still be in the `m-buf`. Not until the message is received and reachable from some other object on the heap or from the stack will the message be copied to the process heap during a GC.

### 3.6.4. Receiving a Message

Erlang supports selective receive, which means that a message that doesn't match can be left in the mailbox for a later receive. And the processes can be suspended with messages in the mailbox when no message matches. The `msg.save` field contains a pointer to a pointer to the next message to look at.

In later chapters we will cover the details of `m-bufs` and how the garbage collector handles mailboxes. We will also go through the details of how receive is implemented in the BEAM in later chapters.

### 3.6.5. Tuning Message Passing

With the new `message_queue_data` flag introduced in Erlang 19 you can trade memory for execution time in a new way. If the receiving process is overloaded and holding on to the `main lock`, it might be a good strategy to use the `off_heap` allocation in order to let the sending process quickly dump the message in an `m-buf`.

If two processes have a nicely balanced producer consumer behavior where there is no real contention for the process lock then allocation directly on the receivers heap will be faster and use less memory.

If the receiver is backed up and is receiving more messages than it has time to handle, it might actually start using more memory as

messages are copied to the heap, and migrated to the old heap. Since unseen messages are considered live, the heap will need to grow and use more memory.

In order to find out which allocation strategy is best for your system you will need to benchmark and measure the behavior. The first and easiest test to do is probably to change the default allocation strategy at the start of the system. The ERTS flag `hmqd` sets the default strategy to either `off_heap` or `on_heap`. If you start Erlang without this flag the default will be `on_heap`. By setting up your benchmark so that Erlang is started with `+hmqd off_heap` you can test whether the system behaves better or worse if all processes use off heap allocation. Then you might want to find bottle neck processes and test switching allocation strategies for those processes only.

## 3.7. The Process Dictionary

There is actually one more memory area in a process where Erlang terms can be stored, the *Process Dictionary*.

The *Process Dictionary* (PD) is a process local key-value store. One advantage with this is that all keys and values are stored on the heap and there is no copying as with send or an ETS table.

We can now update our view of a process with yet another memory area, PD, the process dictionary:

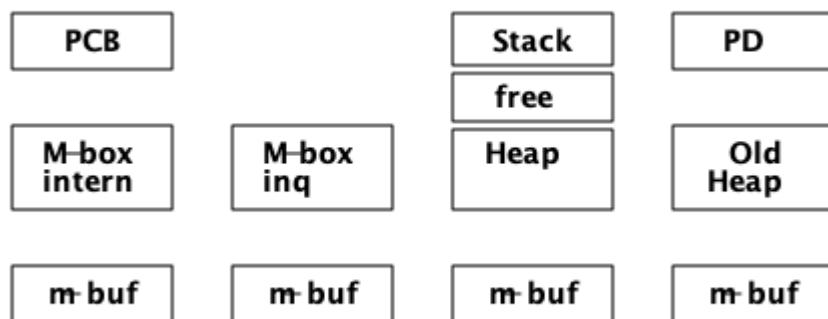


Figure 17. Erlang Process Memory : Process Dictionary

With such a small array you are bound to get some collisions before the area grows. Each hash value points to a bucket with key value pairs. The bucket is actually an Erlang list on the heap. Each entry in the list is a two tuple (`{key, Value}`) also stored on the heap.

Putting an element in the PD is not completely free, it will result in an extra tuple and a cons, and might cause garbage collection to be triggered. Updating a key in the dictionary, which is in a bucket, causes the whole bucket (the whole list) to be reallocated to make sure we don't get pointers from the old heap to the new heap. (In [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#) we will see the details of how garbage collection works.)

## 3.8. Dig In

In this chapter we have looked at how a process is implemented. In particular we looked at how the memory of a process is organized, how message passing works and the information in the PCB. We also looked at a number of tools for inspecting processes introspection, such as `erlang:process_info`, and the `hipe:show*_bifs`.

Use the functions `erlang:processes/0` and `erlang:process_info/1,2` to inspect the processes in the system. Here are some functions to try:

```

1> Ps = erlang:processes().
[<0.0.0>,<0.3.0>,<0.6.0>,<0.7.0>,<0.9.0>,<0.10.0>,<0.11.0>,
 <0.12.0>,<0.13.0>,<0.14.0>,<0.15.0>,<0.16.0>,<0.17.0>,

```

```

<0.19.0>,<0.20.0>,<0.21.0>,<0.22.0>,<0.23.0>,<0.24.0>,
<0.25.0>,<0.26.0>,<0.27.0>,<0.28.0>,<0.29.0>,<0.33.0>]
2> P = self().
<0.33.0>
3> erlang:process_info(P).
[{current_function,{erl_eval,do_apply,6}},
 {initial_call,{erlang,apply,2}},
 {status,running},
 {message_queue_len,0},
 {messages,[]},
 {links,[<0.27.0>]},
 {dictionary,[]},
 {trap_exit,false},
 {error_handler,error_handler},
 {priority,normal},
 {group_leader,<0.26.0>},
 {total_heap_size,17730},
 {heap_size,6772},
 {stack_size,24},
 {reductions,25944},
 {garbage_collection,[{min_bin_vheap_size,46422},
                      {min_heap_size,233},
                      {fullsweep_after,65535},
                      {minor_gcs,1}]},
 {suspending,[]}]
4> lists:ksort(2,[{P,element(2,erlang:process_info(P,
    total_heap_size))} || P <- Ps]).
[{{<0.10.0>,233},
 {<0.13.0>,233},
 {<0.14.0>,233},
 {<0.15.0>,233},
 {<0.16.0>,233},
 {<0.17.0>,233},
 {<0.19.0>,233},
 {<0.20.0>,233},
 {<0.21.0>,233},
 {<0.22.0>,233},
 {<0.23.0>,233},
 {<0.25.0>,233},
 {<0.28.0>,233},
 {<0.29.0>,233},
 {<0.6.0>,752},
 {<0.9.0>,752},
 {<0.11.0>,1363},
 {<0.7.0>,1597},
 {<0.0.0>,1974},
 {<0.24.0>,2585},
 {<0.26.0>,6771},
 {<0.12.0>,13544},
 {<0.33.0>,13544},
 {<0.3.0>,15143},
 {<0.27.0>,32875}]
9>

```



# The Erlang Type System and Tags

One of the most important aspects of ERTS to understand is how ERTS stores data, that is, how Erlang terms are stored in memory. This gives you the basis for understanding how garbage collection works, how message passing works, and gives you an insight into how much memory is needed.

In this chapter you will learn the basic data types of Erlang and how they are implemented in ERTS. This knowledge will be essential in understanding the chapter on memory allocation and garbage collection, see [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#).

## 4.1. The Erlang Type System

Erlang is *strongly typed*. That is, there is no way to coerce one type into another type, you can only convert from one type to another. Compare this to e.g. C where you can coerce a *char* to an *int* or any type pointed to by a pointer to (*void* \*).

The Erlang type lattice is quite flat, there are only a few real sub types, numbers have the sub types integer and float, and list has the subtypes nil and cons. (One could also argue that tuple has one subtype for each size.)

The Erlang Type Lattice

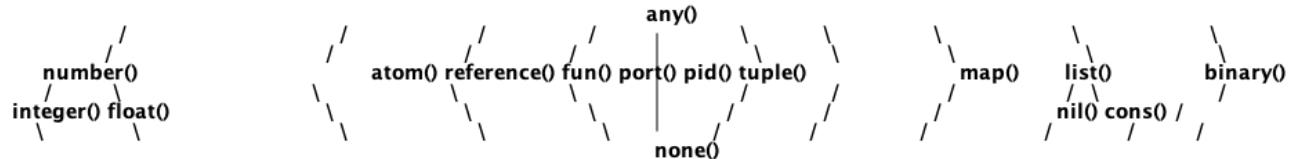


Figure 18. Erlang Type Lattice

There is a partial order (< and >) on all terms in Erlang where the types are ordered from left to right in the above lattice.

The order is partial and not total since integers and floats are converted before comparison. Both  $(1 < 1.0)$  and  $(1.0 < 1)$  are false, and  $(1 = < 1.0 \text{ and } 1 = > 1.0)$  and  $(1 = / 1.0)$ . The number with the lesser precision is converted to the number with higher precision. Usually integers are converted to floats. For very large or small floats the float is converted to an integer. This happens if all significant digits are to the left of the decimal point.

Since Erlang 18, when two maps are compared for order they are compared as follows: If one map has fewer elements than the other it is considered smaller. Otherwise the keys are compared in term order, where all integers are considered smaller than all floats. If all the keys are the same then each value pair (in key order) is compared arithmetically, i.e. by first converting them to the same precision.

The same is true when comparing for equality, thus  $\#\{1 => 1.0\} == \#\{1 => 1\}$  but  $\#\{1.0 => 1\} /= \#\{1 => 1\}$ .

In Erlang versions prior to 18 keys were also compared arithmetically.

Erlang is dynamically typed. That is, types will be checked at runtime and if a type error occurs an exception is thrown. The compiler does not check the types at compile time, unlike in a statically typed language like C or Java where you can get a type error during compilation.

These aspects of the Erlang type system, strongly dynamically typed with an order on the types puts some constraints on the implementation of the language. In order to be able to check and compare types at runtime each Erlang term has to carry its type

with it.

This is solved by *tagging* the terms.

## 4.2. The Tagging Scheme

In the memory representation of an Erlang term a few bits are reserved for a type tag. For performance reasons the terms are divided into *immediates* and *boxed* terms. An immediate term can fit into a machine word, that is, in a register or on a stack slot. A boxed term consists of two parts: a tagged pointer and a number of words stored on the process heap. The *boxes* stored on the heap have a header and a body, unless it is a list.

Currently ERTS uses a staged tag scheme, the history and reasoning behind this scheme is explained in a technical report from the HiPE group. (See <http://www.it.uu.se/research/publications/reports/2000-029/>) The tagging scheme is implemented in erl\_term.h.

The basic idea is to use the least significant bits for tags. Since most modern CPU architectures aligns 32- and 64-bit words, there are at least two bits that are "unused" for pointers. These bits can be used as tags instead. Unfortunately those two bits are not enough for all the types in Erlang, more bits are therefore used as needed.

### 4.2.1. Tags for Immediates

The first two bits (the primary tag) are used as follows:

```
00 Header (on heap) CP (on stack)
01 List (cons)
10 Boxed
11 Immediate
```

The header tag is only used on the heap for header words, more on that later. On the stack 00 indicates a return address. The list tag is used for cons cells, and the boxed tag is used for all other pointers to the heap. The immediate tag is used further divided like this:

```
00 11 Pid
01 11 Port
10 11 Immediate 2
11 11 Small integer
```

Pid and ports are immediates and can be compared for equality efficiently. They are of course in reality just references, a pid is a process identifier and it points to a process. The process does not reside on the heap of any process but is handled by the PCB. A port works in much the same way.

There are two types of integers in ERTS, small integers and bignums. Small integers fits in one machine word minus four tag bits, i.e. in 28 or 60 bits for 32 and 64 bits system respectively. Bignums on the other hand can be as large as needed (only limited by the heap space) and are stored on the heap, as boxed objects.

By having all four tag bits as ones for small integers the emulator can make an efficient test when doing integer arithmetic to see if both arguments are immediates. (is\_both\_small(x,y) is defined as  $(x \& y \& 1111) == 1111$ ).

The Immediate 2 tag is further divided like this:

```
00 10 11 Atom
01 10 11 Catch
10 10 11 [UNUSED]
11 10 11 Nil
```

Atoms are made up of an index in the *atom table* and the atom tag. Two atom immediates can be compared for equality by just comparing their immediate representation.

In the atom table atoms are stored as C structs like this:

```
typedef struct atom {
    IndexSlot slot; /* MUST BE LOCATED AT TOP OF STRUCT!!! */
    int len;          /* length of atom name */
    int ord0;         /* ordinal value of first 3 bytes + 7 bits */
    byte* name;       /* name of atom */
} Atom;
```

Thanks to the len and the ord0 fields the order of two atoms can be compared efficiently as long as they don't start with the same four letters.



If you for some reason generate atoms with a pattern like name followed by a number and then store them in an ordered list or ordered tree the atom comparison will be more expensive if they all have the same first letters (e.g. foo\_1, foo\_2, etc.).

Not that you should ever generate atom names, since the atom table is limited. I'm just saying, there is an evil micro optimization to be found here.

You would of course never do this, but if you find code that generates atom with a number followed by a postfix name, now you know what the author of that code might have been thinking.

The Catch immediate is only used on the stack. It contains an indirect pointer to the continuation point in the code where execution should continue after an exception. More on this in [Different Types of Calls, Linking and Hot Code Loading \(5p\)](#).

The Nil tag is used for the empty list (nil or []). The rest of the word is filled with ones.

### 4.2.2. Tags for Boxed Terms

Erlang terms stored on the heap use several machine words. Lists, or cons cells, are just two consecutive words on the heap: the head and the tail (or car and cdr as they are called in lisp and some places in the ERTS code).

A string in Erlang is just a list of integers representing characters. In releases prior to Erlang OTP R14 strings have been encoded as ISO-latin-1 (ISO8859-1). Since R14 strings are encoded as lists of Unicode code points. For strings in latin-1 there is no difference since latin-1 is a subset of Unicode.

The string "hello" might look like this in memory:

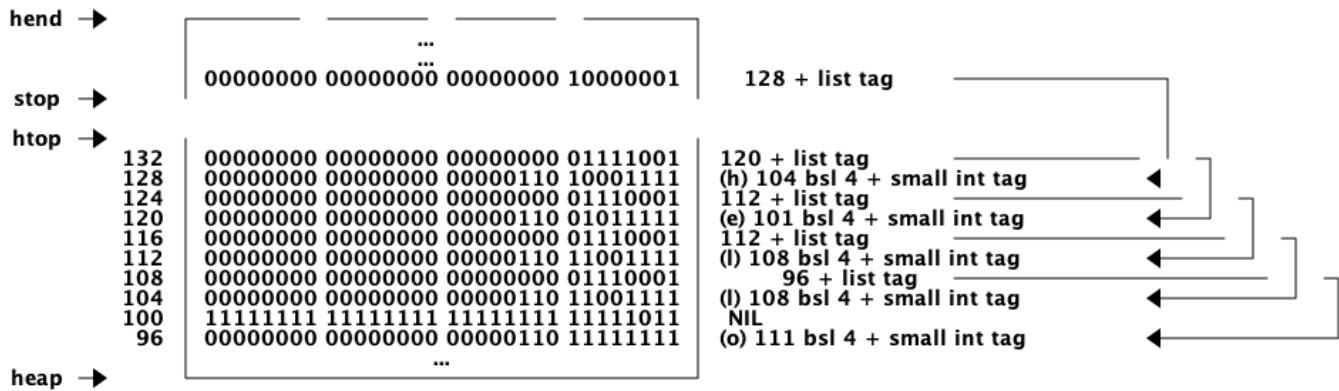


Figure 19. Representation of the string "hello" on a 32 bit machine.

All other boxed terms start with a header word. The header word uses a four bit header tag and the primary header tag (00), it also has an arity which says how many words the boxed term uses. On a 32-bit machine it looks like this: aaaaaaaaaaaaaaaaaaaaaaaaatttt00.

The tags are:

0000	ARITYVAL (Tuples)
0001	BINARY_AGGREGATE
001s	BIGNUM with sign bit
0100	REF
0101	FUN
0110	FLONUM
0111	EXPORT
1000	REFC_BINARY
1001	HEAP_BINARY
1010	SUB_BINARY
1011	[UNUSED]
1100	EXTERNAL_PID
1101	EXTERNAL_PORT
1110	EXTERNAL_REF
1111	MAP
	THINGS
	BINARIES

Tuples are stored on the heap with just the arity and then each element in the following words. The empty tuple {} is stored just as the word 0 (header tag 0, tuple tag 0000, and arity 0).

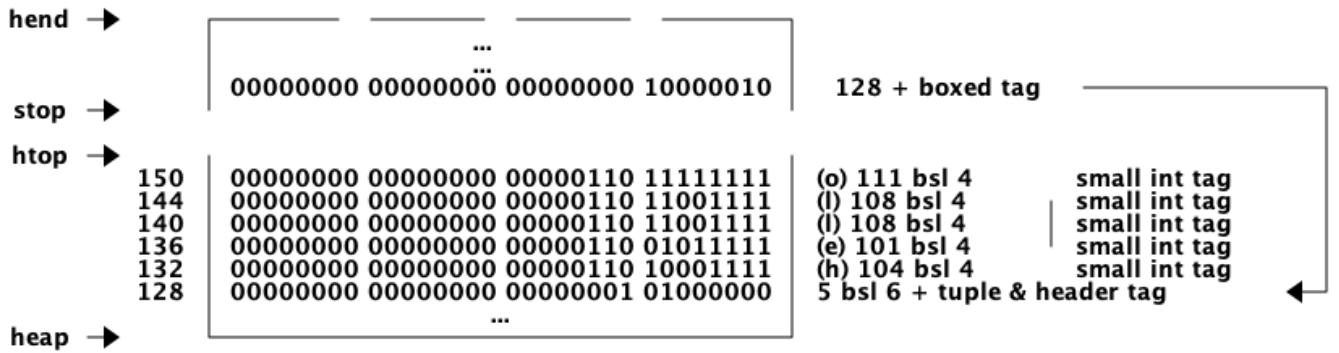


Figure 20. Representation of the tuple {104,101,108,108,111} on a 32 bit machine.

A *binary* is an immutable array of bytes. There are four types of internal representations of *binaries*. The two types *heap binaries* and *refc binaries* contains binary data. The other two types, *sub binaries* and *match contexts* (the `BINARY_AGGREGATE` tag) are smaller references into one of the other two types.

Binaries that are 64 bytes or less can be stored directly on the process heap as *heap binaries*. Larger binaries are reference counted and the payload is stored outside of the process heap, a reference to the payload is stored on the process heap in an object called a *ProcBin*.

We will talk more about binaries in the [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#).

Integers that do not fit in a small integer (word size - 4 bits) are stored on the heap as "bignums" (or arbitrary precision integers). A bignum has a header word followed by a number of words encoding the bignum. The sign part of the bignum tag (s) in the header encodes the sign of the number (s=0 for positive numbers, and s=1 for negative numbers).

TODO: Describe bignum encoding. (And arithmetic ?)

A reference is a "*unique*" term often used to tag messages in order to basically implement a channel over a process mailbox. A reference is implemented as an 82 bit counter. After 967140655691703397649407 calls to `make_ref` the counter will wrap and start over with ref 0 again. You need a really fast machine to do that many calls to `make_ref` within your lifetime. Unless you restart the node, in which case it also will start from 0 again, but then all the old local refs are gone. If you send the pid to another node it becomes an external ref, see below.

On a 32-bit system a local ref takes up four 32-bit words on the heap. On a 64-bit system a ref takes up three 64-bit words on the heap.

*Representation of a ref in a 32-bit (or half-word) system.*

```

|00000000 00000000 00000000 11010000| Arity 3 + ref tag
|00000000 000000rr rrrrrrrr rrrrrrrr| Data0
|rrrrrrrr rrrrrrrr rrrrrrrr rrrrrrrr| Data1
|rrrrrrrr rrrrrrrr rrrrrrrr rrrrrrrr| Data2

```

The reference number is (Data2 bsl 50) + (Data1 bsl 18) + Data0.

## Outline

### TODO

The implementation of floats, ports, pids. Strings as lists, IO lists, lists on 64-bit machines. Binaries, sub binaries, and copying. Records.

Possibly: The half-word machine. Sharing and deep copy. (or this will be in GC)

Outro/conclusion

## Chapter 5

# The Erlang Virtual Machine: BEAM

BEAM (Bogumil's/Björn's Abstract Machine) is the machine that executes the code in the Erlang Runtime System. It is a garbage collecting, reduction counting, virtual, non-preemptive, directly threaded, register machine. If that doesn't tell you much, don't worry, in the following sections we will go through what each of those words means in this context.

The virtual machine, BEAM, is at the heart of the Erlang node. It is the BEAM that executes the Erlang code. That is, it is BEAM that executes your application code. Understanding how BEAM executes the code is vital to be able to profile and tune your code.

The BEAM design influences large parts of the rest of ERTS. The primitives for scheduling influences the Scheduler ([Scheduling](#)), the representation of Erlang terms and the interaction with the memory influences the Garbage Collector ([The Memory Subsystem: Stacks, Heaps and Garbage Collection](#)). By understanding the basic design of BEAM you will more easily understand the implementations of these other components.

## 5.1. Working Memory: A stack machine, it is not

As opposed to its predecessor Jam (Joe's Abstract Machine) which was a stack machine, the BEAM is a register machine loosely based on WAM (TODO: cite). In a stack machine each operand to an instruction is first pushed to the working stack, then the instruction pops its arguments and then it pushes the result on the stack.

Stack machines are quite popular among virtual machine and programming language implementers since they are quite easy to generate code for, and the code becomes very compact. The compiler does not need to do any register allocation, and most operations do not need any arguments (in the instruction stream).

Compiling the expression "8 + 17 \* 2." to a stack machine could yield code like:

```
push 8
push 17
push 2
multiply
add
```

This code can be generated directly from the parse tree of the expression. By using Erlang expression and the modules `erl_scan` and `erl_parse` we can build the world's most simplistic compiler.

```

compile(String) ->
    [ParseTree] = element(2,
        erl_parse:parse_exprs(
            element(2,
                erl_scan:string(String)))),
    generate_code(ParseTree).

generate_code({op, _Line, '+', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [add];
generate_code({op, _Line, '*', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [multiply];
generate_code({integer, _Line, I}) -> [push, I].

```

And an even more simplistic virtual stack machine:

```

interpret(Code) -> interpret(Code, []).

interpret([push, I | Rest], Stack)           -> interpret(Rest, [I|Stack]);
interpret([add      | Rest], [Arg2, Arg1|Stack]) -> interpret(Rest, [Arg1+Arg2|Stack]);
interpret([multiply|Rest], [Arg2, Arg1|Stack]) -> interpret(Rest, [Arg1*Arg2|Stack]);
interpret([],             [Res|_])           -> Res.

```

And a quick test run gives us the answer:

```

1> stack_machine:interpret(stack_machine:compile("8 + 17 * 2.")).
42

```

Great, you have built your first virtual machine! Handling subtraction, division and the rest of the Erlang language is left as an exercise for the reader.

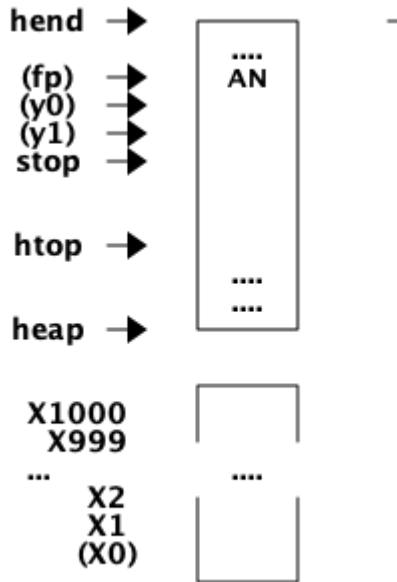
Anyway, the BEAM is **not** a stack machine, it is a *register machine*. In a register machine instruction operands are stored in registers instead of the stack, and the result of an operation usually ends up in a specific register.

Most register machines do still have a stack used for passing arguments to functions and saving return addresses. BEAM has both a stack and registers, but just as in WAM the stack slots are accessible through registers called Y-registers. BEAM also has a number of X-registers, and a special register X0 (sometimes also called R0) which works as an accumulator where results are stored.

The X registers are used as argument registers for function calls and register X0 is used for the return value.

The X registers are stored in a C-array in the BEAM emulator and they are globally accessible from all functions. The X0 register is cached in a local variable mapped to a physical machine register in the native machine on most architectures.

The Y registers are stored in the stack frame of the caller and only accessible by the calling functions. To save a value across a function call BEAM allocates a stack slot for it in the current stack frame and then moves the value to a Y register.



*Figure 21. X and Y Registers in Memory*

Let us compile the following program with the 'S' flag:

```
-module(add).  
-export([add/2]).  
  
add(A,B) -> id(A) + id(B).  
  
id(I) -> I.
```

Then we get the following code for the add function:

```
{function, add, 2, 2}.
```

{label,1}.

```
{func_info,{atom,add},{atom,add},2}.
```

{label,2}.

```
{allocate,1,2}.
```

```
{move,{x,1},{y,0}}.
```

```
{call,1,{f,4}}.
```

```
{move,{x,0},{x,1}}.
```

```
{move,{y,0},{x,0}}.
```

```
{move,{x,1},{y,0}}.
```

```
{call,1,{f,4}}.
```

```
{gc_bif,'+',{f,0},1,[{y,0},{x,0}],{x,0}}.
```

```
{deallocate,1}.
```

```
return.
```

Here we can see that the code (starting at label 2) first allocates a stack slot, to get space to save the argument  $B$  over the function call  $id(A)$ . The value is then saved by the instruction  $\{move,\{x,1\},\{y,0\}\}$  (read as move x1 to y0 or in imperative style:  $y0 := x1$ ).

The id function (at label f4) is then called by  $\{call, I, \{f, 4\}\}$ . (We will come back to what the argument "1" stands for later.) Then the result of the call (now in X0) needs to be saved on the stack (Y0), but the argument B is saved in Y0 so the BEAM does a bit of

shuffling:

Except for the x and y registers, there are a number of special purpose registers:

#### *Special Purpose Registers*

- Htop - The top of the heap.
- E - The top of the stack.
- CP - Continuation Pointer, i.e. function return address
- I - instruction pointer
- fcalls - reduction counter

These registers are cached versions of the corresponding fields in the PCB.

```
{move,{x,0},{x,1}}. % x1 := x0 (id(A))
{move,{y,0},{x,0}}. % x0 := y0 (B)
{move,{x,1},{y,0}}. % y0 := x1 (id(A))
```

Now we have the second argument *B* in x0 (the first argument register) and we can call the *id* function again *{call,1,{f,4}}*.

After the call x0 contains *id(B)* and y0 contains *id(A)*, now we can do the addition: *{gc\_bif,'+',{f,0},1,[{y,0},{x,0}],{x,0}}*. (We will go into the details of BIF calls and GC later.)

## 5.2. Dispatch: Directly Threaded Code

The instruction decoder in BEAM is implemented with a technique called *directly threaded* code. In this context the word *threaded* has nothing to do with OS threads, concurrency or parallelism. It is the execution path which is threaded through the virtual machine itself.

If we take a look at our naive stack machine for arithmetic expressions we see that we use Erlang atoms and pattern matching to decode which instruction to execute. This is a very heavy machinery to just decode machine instructions. In a real machine we would code each instruction as a "machine word" integer.

We can rewrite our stack machine to be a *byte code* machine implemented in C. First we rewrite the compiler so that it produces byte codes. This is pretty straight forward, just replace each instruction encoded as an atom with a byte representing the instruction. To be able to handle integers larger than 255 we encode integers with a size byte followed by the integer encoded in bytes.

```

compile(Expression, FileName) ->
    [ParseTree] = element(2,
        erl_parse:parse_exprs(
            element(2,
                erl_scan:string(Expression)))),
    file:write_file(Filename, generate_code(ParseTree) ++ [stop()]). 

generate_code({op, _Line, '+', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [add()];
generate_code({op, _Line, '*', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [multiply()];
generate_code({integer, _Line, I}) -> [push(), integer(I)]. 

stop()      -> 0.
add()       -> 1.
multiply()  -> 2.
push()      -> 3.
integer(I)  ->
    L = binary_to_list(binary:encode_unsigned(I)),
    [length(L) | L].

```

Now lets write a simple virtual machine in C. The full code can be found in [Full Code Listings](#).

```

#define STOP 0
#define ADD 1
#define MUL 2
#define PUSH 3

#define pop() (stack[--sp])
#define push(X) (stack[sp++] = X)

int run(char *code) {
    int stack[1000];
    int sp = 0, size = 0, val = 0;
    char *ip = code;

    while (*ip != STOP) {
        switch (*ip++) {
        case ADD: push(pop() + pop()); break;
        case MUL: push(pop() * pop()); break;
        case PUSH:
            size = *ip++;
            val = 0;
            while (size--) { val = val * 256 + *ip++; }
            push(val);
            break;
        }
    }
    return pop();
}

```

You see, a virtual machine written in C does not need to be very complicated. This machine is just a loop checking the byte code at

each instruction by looking at the value pointed to by the \_instruction pointer \_ (ip).

For each byte code instruction it will switch on the instruction byte code and jump to the case which executes the instruction. This requires a decoding of the instruction and then a jump to the correct code. If we look at the assembly for vsm.c (gcc -S vsm.c) we see the inner loop of the decoder:

L11:

```
    movl    -16(%ebp), %eax
    movzbl (%eax), %eax
    movsbl %al, %eax
    addl    $1, -16(%ebp)
    cmpl    $2, %eax
    je     L7
    cmpl    $3, %eax
    je     L8
    cmpl    $1, %eax
    jne    L5
```

It has to compare the byte code with each instruction code and then do a conditional jump. In a real machine with many instructions this can become quite expensive.

A better solution would be to have a table with the address of the code then we could just use an index into the table to load the address and jump without the need to do a compare. This technique is sometimes called *token threaded code*. Taking this a step further we can actually store the address of the function implementing the instruction in the code memory. This is called *subroutine threaded code*.

This approach will make the decoding simpler at runtime, but it makes the whole VM more complicated by requiring a loader. The loader replaces the byte code instructions with addresses to functions implementing the instructions.

A loader might look like:

```

typedef void (*instructionp_t)(void);

instructionp_t *read_file(char *name) {
    FILE *file;
    instructionp_t *code;
    instructionp_t *cp;
    long size;
    char ch;
    unsigned int val;

    file = fopen(name, "r");
    if(file == NULL) exit(1);

    fseek(file, 0L, SEEK_END);
    size = ftell(file);
    code = calloc(size, sizeof(instructionp_t));
    if(code == NULL) exit(1);
    cp = code;

    fseek(file, 0L, SEEK_SET);
    while ( ( ch = fgetc(file) ) != EOF )
    {
        switch (ch) {
            case ADD: *cp++ = &add; break;
            case MUL: *cp++ = &mul; break;
            case PUSH:
                *cp++ = &pushi;
                ch = fgetc(file);
                val = 0;
                while (ch--) { val = val * 256 + fgetc(file); }
                *cp++ = (instructionp_t) val;
                break;
        }
    }
    *cp = &stop;

    fclose(file);
    return code;
}

```

As we can see, we do more work at load time here, including the decoding of integers larger than 255. (Yes, I know, the code is not safe for very large integers.)

The decode and dispatch loop of the VM becomes quite simple though:

```

int run() {
    sp = 0;
    running = 1;

    while (running) (*ip++)();

    return pop();
}

```

Then we just need to implement the instructions:

```

void add() { int x,y; x = pop(); y = pop(); push(x + y); }
void mul() { int x,y; x = pop(); y = pop(); push(x * y); }
void pushi(){ int x; x = (int)*ip++; push(x); }
void stop() { running = 0; }

```

In BEAM this concept is taken one step further, and BEAM uses *directly threaded code* (sometimes called only *thread code*). In directly threaded code the call and return sequence is replaced by direct jumps to the implementation of the next instruction. In order to implement this in C, BEAM uses the GCC extension "labels as values".

We will look closer at the BEAM emulator later but we will take a quick look at how the add instruction is implemented. The code is somewhat hard to follow due to the heavy usage of macros. The STORE\_ARITH\_RESULT macro actually hides the dispatch function which looks something like: `I += 4; Goto(*I);`.

```

#define OpCase(OpCode)    lb_##OpCode
#define Goto(Rel) goto *(Rel)

...
OpCase(i_plus_jId):
{
    Eterm result;

    if (is_both_small(tmp_arg1, tmp_arg2)) {
        Sint i = signed_val(tmp_arg1) + signed_val(tmp_arg2);
        ASSERT(MY_IS_SSMALL(i) == IS_SSMALL(i));
        if (MY_IS_SSMALL(i)) {
            result = make_small(i);
            STORE_ARITH_RESULT(result);
        }
    }
    arith_func = ARITH_FUNC(mixed_plus);
    goto do_big_arith2;
}

```

To make it a little easier to understand how the BEAM dispatcher is implemented let us take a somewhat imaginary example. We will start with some real external BEAM code but then I will invent some internal BEAM instructions and implement them in C.

If we start with a simple add function in Erlang:

```
add(A,B) -> id(A) + id(B).
```

Compiled to BEAM code this will look like:

```
{function, add, 2, 2}.
{label,1}.
{func_info,{atom,add},{atom,add},2}.
{label,2}.
{allocate,1,2}.
{move,{x,1},{y,0}}.
{call,1,{f,4}}.
{move,{x,0},{x,1}}.
{move,{y,0},{x,0}}.
{move,{x,1},{y,0}}.
{call,1,{f,4}}.
{gc_bif,'+',{f,0},1,[{y,0},{x,0}],{x,0}}.
{deallocate,1}.
{return}.
```

(See add.erl and add.S in [Full Code Listings](#) for the full code.)

Now if we zoom in on the three instructions between the function calls in this code:

```
{move,{x,0},{x,1}}.
{move,{y,0},{x,0}}.
{move,{x,1},{y,0}}.
```

This code first saves the return value of the function call (x0) in a new register (x1). Then it moves the caller saves register (y0) to the first argument register (x0). Finally it moves the saved value in x1 to the caller save register (y0) so that it will survive the next function call.

Imagine that we would implement three instruction in BEAM called move\_xx, move\_yx, and move\_xy (These instructions does not exist in the BEAM we just use them to illustrate this example):

```
#define OpCase(OpCode)    lb_##OpCode
#define Goto(Rel) goto *((void *)Rel)
#define Arg(N) (Eterm *) I[(N)+1]
```

```
OpCase(move_xx):
{
    x(Arg(1)) = x(Arg(0));
    I += 3;
    Goto(*I);
}
```

```
OpCase(move_yx): {
    x(Arg(1)) = y(Arg(0));
    I += 3;
    Goto(*I);
}
```

```
OpCase(move_xy): {
    y(Arg(1)) = x(Arg(0));
    I += 3;
    Goto(*I);
}
```

Note that the star in `goto *` does not mean dereference, the expression means jump to an address pointer, we should really write it as `goto*`.

Now imagine that the compiled C code for these instructions end up at memory addresses 0x3000, 0x3100, and 0x3200. When the BEAM code is loaded the three move instructions in the code will be replaced by the memory addresses of the implementation of the instructions. Imagine that the code (`{move,{x,0},{x,1}}`, `{move,{y,0},{x,0}}`, `{move,{x,1},{y,0}}`) is loaded at address 0x1000:

```
/ 0x1000: 0x3000 -> 0x3000: OpCase(move_xx): x(Arg(1)) = x(Arg(0))
{move,{x,0},{x,1}} { 0x1004: 0x0                         I += 3;
                      \ 0x1008: 0x1                         Goto(*I);
                      / 0x100c: 0x3100
{move,{y,0},{x,0}} { 0x1010: 0x0
                      \ 0x1014: 0x0
                      / 0x1018: 0x3200
{move,{x,1},{y,0}} { 0x101c: 0x1
                      \ 0x1020: 0x0
```

The word at address 0x1000 points to the implementation of the `move_xx` instruction. If the register `I` contains the instruction pointer, pointing to 0x1000 then the dispatch will be to fetch `*I` (i.e. 0x3000) and jump to that address. (`goto* *I`)

In [Generic BEAM Instructions \(25p\)](#) we will look more closely at some real BEAM instructions and how they are implemented.

## 5.3. Scheduling: Non-preemptive, Reduction counting

Most modern multi-threading operating systems use preemptive scheduling. This means that the operating system decides when to switch from one process to another, regardless of what the process is doing. This protects the other processes from a process

misbehaving by not yielding in time.

In cooperative multitasking which uses a non-preemptive scheduler the running process decides when to yield. This has the advantage that the yielding process can do so in a known state.

For example in a language such as Erlang with dynamic memory management and tagged values, an implementation may be designed such that a process only yields when there are no untagged values in working memory.

Take the add instruction as an example, to add two Erlang integers, the emulator first has to untag the integers, then add them together and then tag the result as an integer. If a fully preemptive scheduler is used there would be no guarantee that the process isn't suspended while the integers are untagged. Or the process could be suspended while it is creating a tuple on the heap, leaving us with half a tuple. This would make it very hard to traverse a suspended process stack and heap.

On the language level all processes are running concurrently and the programmer should not have to deal with explicit yields. BEAM solves this by keeping track of how long a process has been running. This is done by counting *reductions*. The term originally comes from the mathematical term beta-reduction used in lambda calculus.

The definition of a reduction in BEAM is not very specific, but we can see it as a small piece of work, which shouldn't take *too long*. Each function call is counted as a reduction. BEAM does a test upon entry to each function to check whether the process has used up all its reductions or not. If there are reductions left the function is executed otherwise the process is suspended.

Since there are no loops in Erlang, only tail-recursive function calls, it is very hard to write a program that does any significant amount of work without using up its reductions.

 There are some BIFs that can run for a long time only using 1 reduction, like `term_to_binary` and `binary_to_term`. Try to make sure that you only call these BIFs with small terms or binaries, or you might lock up the scheduler for a very long time.

Also, if you write your own NIFs, make sure they can yield and that they bump the reduction counter by an amount proportional to their run time.

We will go through the details of how the scheduler works in [Scheduling](#).

## 5.4. Memory Management: Garbage Collecting

Erlang supports garbage collection; as an Erlang programmer you do not need to do explicit memory management. On the BEAM level, though, the code is responsible for checking for stack and heap overrun, and for allocating enough space on the stack and the heap.

The BEAM instruction `test_heap` will ensure that there is as much space on the heap as requested. If needed the instruction will call the garbage collector to reclaim space on the heap. The garbage collector in turn will call the lower levels of the memory subsystem to allocate or free memory as needed. We will look at the details of memory management and garbage collection in [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#).

## 5.5. BEAM: it is virtually unreal

The BEAM is a virtual machine, by that we mean that it is implemented in software instead of in hardware. There has been projects to implement the BEAM by FPGA, and there is nothing stopping anyone from implementing the BEAM in hardware. A better description might be to call the BEAM an Abstract machine, and see it as blueprint for a machine which can execute BEAM code. And, in fact, the "am" in BEAM stands for "Abstract Machine".

In this book we will make no distinction between abstract machines, and virtual machines or their implementation. In a more formal setting an abstract machine is a theoretical model of a computer, and a virtual machine is either a software implementation of an abstract machine or a software emulator of a real physical machine.

Unfortunately there exist no official specification of the BEAM, it is currently only defined by the implementation in Erlang/OTP. If you want to implement your own BEAM you would have to try to mimic the current implementation not knowing which parts are essential and which parts are accidental. You would have to mimic every observable behavior to be sure that you have a valid

BEAM interpreter.

**TODO:** Conclusion and handover to the chapters on instructions.

## Chapter 6

# Modules and The BEAM File Format (16p)

## 6.1. Modules

*%% TODO*

 What is a module. How is code loaded. How does hot code loading work. How does the purging work. How does the code server work. How does dynamic code loading work, the code search path. Handling code in a distributed system. (Overlap with chapter 10, have to see what goes where.) Parameterized modules. How p-mods are implemented. The trick with p-mod calls. Here follows an excerpt from the current draft:

## 6.2. The BEAM File Format

The definite source of information about the beam file format is obviously the source code of beam\_lib.erl (see [https://github.com/erlang/otp/blob/maint/lib/stdlib/src/beam\\_lib.erl](https://github.com/erlang/otp/blob/maint/lib/stdlib/src/beam_lib.erl)). There is actually also a more readable but slightly dated description of the format written by the main developer and maintainer of Beam (see [http://www.erlang.se/~bjorn/beam\\_file\\_format.html](http://www.erlang.se/~bjorn/beam_file_format.html)).

The beam file format is based on the interchange file format (EA IFF) #, with two small changes. We will get to those shortly. An IFF file starts with a header followed by a number of “chunks”. There are a number of standard chunk types in the IFF specification dealing mainly with images and music. But the IFF standard also lets you specify your own named chunks, and this is what BEAM does.

 Beam files differ from standard IFF files, in that each chunk is aligned on 4-byte boundary (i.e. 32 bit word) instead of on a 2-byte boundary as in the IFF standard. To indicate that this is not a standard IFF file the IFF header is tagged with “FOR1” instead of “FOR”. The IFF specification suggests this tag for future extensions.

Beam uses form type “BEAM”. A beam file header has the following layout:

```
BEAMHeader = <<
  IffHeader:4/unit:8 = "FOR1",
  Size:32/big,           // big endian, how many more bytes are there
  FormType:4/unit:8 = "BEAM"
>>
```

After the header multiple chunks can be found. Size of each chunk is aligned to the multiple of 4 and each chunk has its own header (below).

 The alignment is important for some platforms, where unaligned memory byte access would create a hardware exception (named SIGBUS in Linux). This can turn out to be a performance hit or the exception could crash the VM.

```

BEAMChunk = <<
    ChunkName:4/unit:8,           // "Code", "Atom", "StrT", "LitT", ...
    ChunkSize:32/big,
    ChunkData:ChunkSize/unit:8,   // data format is defined by ChunkName
    Padding4:0..3/unit:8
>>

```

This file format prepends all areas with the size of the following area making it easy to parse the file directly while reading it from disk. To illustrate the structure and content of beam files, we will write a small program that extract all the chunks from a beam file. To make this program as simple and readable as possible we will not parse the file while reading, instead we load the whole file in memory as a binary, and then parse each chunk. The first step is to get a list of all chunks:

```

<!-- While this is reasonably simple code, is it going to be obvious to readers what's going on here?
It may require a bit of explanation as to how this code works. - bmacdonald -->

```

```

-module(beamfile).
-export([read/1]).

read(Filename) ->
{ok, File} = file:read_file(Filename),
<<"FOR1",
Size:32/integer,
"BEAM",
Chunks/binary>> = File,
{Size, read_chunks(Chunks, [])}.

read_chunks(<<N,A,M,E, Size:32/integer, Tail/binary>>, Acc) ->
% Align each chunk on even 4 bytes
ChunkLength = align_by_four(Size),
<<Chunk:ChunkLength/binary, Rest/binary>> = Tail,
read_chunks(Rest, [{[N,A,M,E], Size, Chunk}|Acc]);
read_chunks(<<>>, Acc) -> lists:reverse(Acc).

align_by_four(N) -> (4 * ((N+3) div 4)).

```

A sample run might look like:

```
> beamfile:read("beamfile.beam").
{848,
 [{"Atom",103,
   <<0,0,0,14,4,102,111,114,49,4,114,101,97,100,4,102,105,
    108,101,9,114,101,97,...>>},
 {"Code",341,
   <<0,0,0,16,0,0,0,0,0,0,0,132,0,0,0,14,0,0,0,4,1,16,...>>},
 {"StrT",8,<<"FOR1BEAM">>},
 {"ImpT",88,<<0,0,0,7,0,0,0,3,0,0,0,4,0,0,0,1,0,0,0,7,...>>},
 {"ExpT",40,<<0,0,0,3,0,0,0,13,0,0,0,1,0,0,0,13,0,0,0,...>>},
 {"LocT",16,<<0,0,0,1,0,0,0,6,0,0,0,2,0,0,0,6>>},
 {"Attr",40,
   <<131,108,0,0,0,1,104,2,100,0,3,118,115,110,108,0,0,...>>},
 {"CInf",130,
   <<131,108,0,0,0,4,104,2,100,0,7,111,112,116,105,111,...>>},
 {"Abst",0,<<>>}]}
```

Here we can see the chunk names that beam uses.

### 6.2.1. Atom table chunk

Either the chunk named **Atom** or the chunk named **AtU8** is mandatory. It contains all atoms referred to by the module. For source files with **latin1** encoding, the chunk named **Atom** is used. For **utf8** encoded modules, the chunk is named **AtU8**. The format of the atom chunk is:

```
AtomChunk = <<
  ChunkName:4/unit:8 = "Atom",
  ChunkSize:32/big,
  NumberOfAtoms:32/big,
  [<<AtomLength:8, AtomName:AtomLength/unit:8>> || repeat NumberOfAtoms],
  Padding4:0..3/unit:8
>>
```

The format of the AtU8 chunk is the same as above, except that the name of the chunk is **AtU8**.



Module name is always stored as the first atom in the table (atom index 0).

Let us add a decoder for the atom chunk to our Beam file reader:

```

-module(beanfile).
-export([read/1]).

read(Filename) ->
{ok, File} = file:read_file(Filename),
<<"FOR1",
  Size:32/integer,
  "BEAM",
  Chunks/binary>> = File,
{Size, parse_chunks(read_chunks(Chunks, []), [])}.

read_chunks(<<N,A,M,E, Size:32/integer, Tail/binary>>, Acc) ->
% Align each chunk on even 4 bytes
ChunkLength = align_by_four(Size),
<<Chunk:ChunkLength/binary, Rest/binary>> = Tail,
read_chunks(Rest, [{[N,A,M,E], Size, Chunk}|Acc]);
read_chunks(<<>>, Acc) -> lists:reverse(Acc).

parse_chunks([{Atom, _Size,
  <<_Numberofatoms:32/integer, Atoms/binary>>}
 | Rest], Acc) ->
  parse_chunks(Rest, [{atoms, parse_atoms(Atoms)}|Acc]);
parse_chunks([Chunk|Rest], Acc) -> %% Not yet implemented chunk
  parse_chunks(Rest, [Chunk|Acc]);
parse_chunks([], Acc) -> Acc.

parse_atoms(<<Atomlength, Atom:Atomlength/binary, Rest/binary>>) when Atomlength > 0->
  [list_to_atom(binary_to_list(Atom)) | parse_atoms(Rest)];
parse_atoms(_Alignment) -> [].

align_by_four(N) -> (4 * ((N+3) div 4)).

```

## 6.2.2. Export table chunk

The chunk named `ExpT` (for EXPort Table) is mandatory and contains information about which functions are exported.

The format of the export chunk is:

```

ExportChunk = <<
  ChunkName:4/unit:8 = "ExpT",
  ChunkSize:32/big,
  ExportCount:32/big,
  [ << FunctionName:32/big,
    Arity:32/big,
    Label:32/big
    >> || repeat ExportCount ],
  Padding4:0..3/unit:8
>>

```

`FunctionName` is the index in the atom table.

We can extend our `parse_chunk` function by adding the following clause after the atom handling clause:

```

parse_chunks([{"ExpT", _Size,
    <<_Numberofentries:32/integer, Exports/binary>>}
 | Rest], Acc) ->
parse_chunks(Rest,[{exports,parse_exports(Exports)}|Acc]);

...

parse_exports(<<Function:32/integer,
    Arity:32/integer,
    Label:32/integer,
    Rest/binary>>) ->
[{Function, Arity, Label} | parse_exports(Rest)];
parse_exports(<<>>) -> [].

```

### 6.2.3. Import table chunk

The chunk named **ImpT** (for IMPort Table) is mandatory and contains information about which functions are imported.

The format of the chunk is:

```

ImportChunk = <<
    ChunkName:4/unit:8 = "ImpT",
    ChunkSize:32/big,
    ImportCount:32/big,
    [ << ModuleName:32/big,
        FunctionName:32/big,
        Arity:32/big
      >> || repeat ImportCount ],
    Padding4:0..3/unit:8
  >>

```

Here **ModuleName** and **FunctionName** are indexes in the atom table.



The code for parsing the import table is similar to that which parses the export table, but not exactly: both are triplets of 32-bit integers, just their meaning is different. See the full code at the end of the chapter.

### 6.2.4. Code Chunk

The chunk named **Code** contains the beam code for the module and is mandatory. The format of the chunk is:

```

ImportChunk = <<
    ChunkName:4/unit:8 = "Code",
    ChunkSize:32/big,
    SubSize:32/big,
    InstructionSet:32/big,           % Must match code version in the emulator
    OpcodeMax:32/big,
    LabelCount:32/big,
    FunctionCount:32/big,
    Code:(ChunkSize-SubSize)/binary, % all remaining data
    Padding4:0..3/unit:8
>>

```

The field **SubSize** stores the number of words before the code starts. This makes it possible to add new information fields in the code chunk without breaking older loaders.

The **InstructionSet** field indicates which version of the instruction set the file uses. The version number is increased if any instruction is changed in an incompatible way.

The **OpcodeMax** field indicates the highest number of any opcode used in the code. New instructions can be added to the system in a way such that older loaders still can load a newer file as long as the instructions used in the file are within the range the loader knows about.

The field **LabelCount** contains the number of labels so that a loader can preallocate a label table of the right size in one call. The field **FunctionCount** contains the number of functions so that the functions table could also be preallocated efficiently.

The **Code** field contains instructions, chained together, where each instruction has the following format:

```

Instruction = <<
    InstructionCode:8,
    [beam_asm:encode(Argument) || repeat Arity]
>>

```

Here **Arity** is hardcoded in the table, which is generated from ops.tab by genop script when the emulator is built from source.

The encoding produced by **beam\_asm:encode** is explained below in the [Compact Term Encoding](#) section.

We can parse out the code chunk by adding the following code to our program:

```

parse_chunks([{{"Code", Size, <<SubSize:32/integer,Chunk/binary>>
    } | Rest], Acc) ->
<<Info:SubSize/binary, Code/binary>> = Chunk,
% 8 is size of CunkSize & SubSize
OpcodeSize = Size - SubSize - 8,
<<OpCodes:OpcodeSize/binary, _Align/binary>> = Code,
parse_chunks(Rest,[{code,parse_code_info(Info), OpCodes}
    | Acc]);
.

.

parse_code_info(<<Instructionset:32/integer,
    OpcodeMax:32/integer,
    NumberOfLabels:32/integer,
    NumberOfFunctions:32/integer,
    Rest/binary>>) ->
[{instructionset, Instructionset},
 {opcodemax, OpcodeMax},
 {numberoflabels, NumberOfLabels},
 {numberoffunctions, NumberOfFunctions} |
 case Rest of
 <<>> -> [];
 _ -> [{newinfo, Rest}]
end].

```

We will learn how to decode the beam instructions in a later chapter, aptly named “BEAM Instructions”.

### 6.2.5. String table chunk

The chunk named `StrT` is mandatory and contains all constant string literals in the module as one long string. If there are no string literals the chunks should still be present but empty and of size 0.

The format of the chunk is:

```

StringChunk = <<
    ChunkName:4/unit:8 = "StrT",
    ChunkSize:32/big,
    Data:ChunkSize/binary,
    Padding4:0..3/unit:8
>>

```

The string chunk can be parsed easily by just turning the string of bytes into a binary:

```

parse_chunks([{{"StrT", _Size, <<Strings/binary>>} | Rest], Acc) ->
    parse_chunks(Rest,[{strings,binary_to_list(Strings)}|Acc]);

```

### 6.2.6. Attributes Chunk

The chunk named `Attr` is optional, but some OTP tools expect the attributes to be present. The release handler expects the “vsn” attribute to be present. You can get the version attribute from a file with: `beam_lib:version(Filename)`, this function assumes that

there is an attribute chunk with a "vsn" attribute present.

The format of the chunk is:

```
AttributesChunk = <<
  ChunkName:4/unit:8 = "Attr",
  ChunkSize:32/big,
  Attributes:ChunkSize/binary,
  Padding4:0..3/unit:8
>>
```

We can parse the attribute chunk like this:

```
parse_chunks([{{"Attr", Size, Chunk} | Rest}, Acc] ->
  <<Bin:Size/binary, _Pad/binary>> = Chunk,
  Attribs = binary_to_term(Bin),
  parse_chunks(Rest,[{attributes,Attribs}|Acc]);
```

## 6.2.7. Compilation Information Chunk

The chunk named **CInf** is optional, but some OTP tools expect the information to be present.

The format of the chunk is:

```
CompilationInfoChunk = <<
  ChunkName:4/unit:8 = "CInf",
  ChunkSize:32/big,
  Data:ChunkSize/binary,
  Padding4:0..3/unit:8
>>
```

We can parse the compilation information chunk like this:

```
parse_chunks([{{"CInf", Size, Chunk} | Rest}, Acc] ->
  <<Bin:Size/binary, _Pad/binary>> = Chunk,
  CInfo = binary_to_term(Bin),
  parse_chunks(Rest,[{compile_info,CInfo}|Acc]);
```

## 6.2.8. Local Function Table Chunk

The chunk named **LocT** is optional and intended for cross reference tools.

The format is the same as that of the export table:

```

LocalFunTableChunk = <<
  ChunkName:4/unit:8 = "LocT",
  ChunkSize:32/big,
  FunctionCount:32/big,
  [ << FunctionName:32/big,
    Arity:32/big,
    Label:32/big
    >> || repeat FunctionCount ],
  Padding4:0..3/unit:8
>>

```



The code for parsing the local function table is basically the same as that for parsing the export and the import table, and we can actually use the same function to parse entries in all tables. See the full code at the end of the chapter.

### 6.2.9. Literal Table Chunk

The chunk named `LitT` is optional and contains all literal values from the module source in compressed form, which are not immediate values. The format of the chunk is:

```

LiteralTableChunk = <<
  ChunkName:4/unit:8 = "LitT",
  ChunkSize:32/big,
  UncompressedSize:32/big,      % It is nice to know the size to allocate some memory
  CompressedLiterals:ChunkSize/binary,
  Padding4:0..3/unit:8
>>

```

Where the `CompressedLiterals` must have exactly `UncompressedSize` bytes. Each literal in the table is encoded with the External Term Format (`erlang:term_to_binary`). The format of `CompressedLiterals` is the following:

```

CompressedLiterals = <<
  Count:32/big,
  [ <<Size:32/big, Literal:binary>> || repeat Count ]
>>

```

The whole table is compressed with `zlib:compress/1` (deflate algorithm), and can be uncompressed with `zlib:uncompress/1` (inflate algorithm).

We can parse the chunk like this:

```

parse_chunks([{"LitT", _ChunkSize,
              <<_CompressedTableSize:32, Compressed/binary>>}
             | Rest], Acc) ->
  <<_NumLiterals:32, Table/binary>> = zlib:uncompress(Compressed),
  Literals = parse_literals(Table),
  parse_chunks(Rest,[{literals,Literals}|Acc]);

```

...

```
parse_literals(<<Size:32,Literal:Size/binary,Tail/binary>>) ->
    [binary_to_term(Literal) | parse_literals(Tail)];
parse_literals(<<>>) -> [].
```

## 6.2.10. Abstract Code Chunk

The chunk named `Abst` is optional and may contain the code in abstract form. If you give the flag `debug_info` to the compiler it will store the abstract syntax tree for the module in this chunk. OTP tools like the debugger and Xref need the abstract form. The format of the chunk is:

```
AbstractCodeChunk = <<
    ChunkName:4/unit:8 = "Abst",
    ChunkSize:32/big,
    AbstractCode:ChunkSize/binary,
    Padding4:0..3/unit:8
>>
```

We can parse the chunk like this

```
parse_chunks([{{"Abst", _ChunkSize, <<>>} | Rest], Acc) ->
    parse_chunks(Rest, Acc);
parse_chunks([{{"Abst", _ChunkSize, <<AbstractCode/binary>>} | Rest], Acc) ->
    parse_chunks(Rest, [{abstract_code,binary_to_term(AbstractCode)}|Acc]);
```

## 6.2.11. Compression and Encryption

TODO

## 6.2.12. Function Trace Chunk (Obsolete)

This chunk type is now obsolete.

## 6.2.13. Bringing it all Together

TODO

## 6.2.14. Compact Term Encoding

Let's look at the algorithm, used by `beam_asm:encode`. BEAM files use a special encoding to store simple terms in BEAM file in a space-efficient way. It is different from memory term layout, used by the VM.



`Beam_asm` is a module in the `compiler` application, part of the Erlang distribution, it is used to assemble binary content of beam modules.

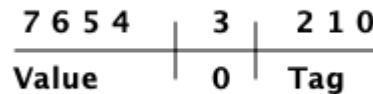
The reason behind this complicated design is to try and fit as much type and value data into the first byte as possible to make the code section more compact. After decoding, all encoded values become full size machine words or terms.

7 6 5 4 3	2 1 0		
0 0 0	0 0 0	Literal	(tag_u in beam_opcodes.hrl)
0 0 1	0 0 1	Integer	(tag_i)
0 1 0	0 1 0	Atom	(tag_a)
0 1 1	0 1 1	X Register	(tag_x)
1 0 0	1 0 0	Y Register	(tag_y)
1 0 1	1 0 1	Label	(tag_f)
0 0 0 1 0	1 1 0	Character	(tag_h)
0 0 1 0 0	1 1 1	Extended	– Float
0 0 1 1 0	1 1 1	Extended	– List
0 1 0 0 0	1 1 1	Extended	– Floating point register
0 1 0 1 0	1 1 1	Extended	– Allocation list
		Extended	– Literal

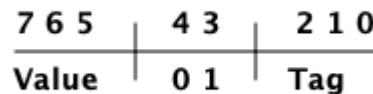
 Since OTP 20 this tag format has been changed and the **Extended - Float** is gone. All following tag values are shifted down by 1: List is 2#10111, freg is 2#100111, alloc list is 2#110111 and literal is 2#1010111. Floating point values now go straight to literal area of the BEAM file.

It uses the first 3 bits of a first byte to store the tag which defines the type of the following value. If the bits were all 1 (special value 7 or ?tag\_z from `beam_opcodes.hrl`), then a few more bits are used.

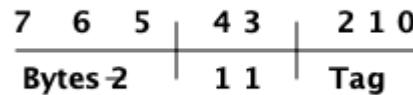
For values under 16, the value is placed entirely into bits 4-5-6-7 having bit 3 set to 0:



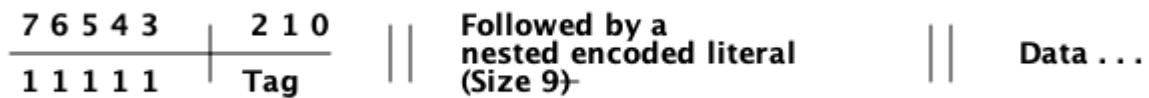
For values under 2048 (16#800) bit 3 is set to 1, marks that 1 continuation byte will be used and 3 most significant bits of the value will extend into this byte's bits 5-6-7:



Larger and negative values are first converted to bytes. Then if the value takes 2..8 bytes, bits 3-4 will be set to 1, and bits 5-6-7 will contain the (Bytes-2) size for the value, which follows:



If the following value is greater than 8 bytes, then all bits 3-4-5-6-7 will be set to 1, followed by a nested encoded unsigned literal (macro `?tag_u` in `beam_opcodes.hrl`) value of (Bytes-9):8, and then the data:



## Tag Types

When reading compact term format, the resulting integer may be interpreted differently based on what is the value of **Tag**.

- For literals the value is index into the literal table.
- For atoms, the value is atom index MINUS one. If the value is 0, it means **NIL** (empty list) instead.
- For labels 0 means invalid value.
- If tag is character, the value is unsigned unicode codepoint.
- Tag Extended List contains pairs of terms. Read **Size**, create tuple of **Size** and then read **Size/2** pairs into it. Each pair is **Value** and **Label**. **Value** is a term to compare against and **Label** is where to jump on match. This is used in **select\_val** instruction.

Refer to [beam\\_asm:encode/2](#) in the compiler application for details about how this is encoded. Tag values are presented in this section, but also can be found in [compiler/src/beam\\_opcodes.hrl](#).

**Chapter 7****Generic BEAM Instructions (25p)**

Beam has two different instructions sets, an internal instructions set, called *specific*, and an external instruction set, called *generic*.

The generic instruction set is what could be called the official instruction set, this is the set of instructions used by both the compiler and the Beam interpreter. If there was an official Erlang Virtual Machine specification it would specify this instruction set. If you want to write your own compiler to the Beam, this is the instruction set you should target. If you want to write your own EVM this is the instruction set you should handle.

The external instruction set is quite stable, but it does change between Erlang versions, especially between major versions.

This is the instruction set which we will cover in this chapter.

The other instruction set, the specific, is an optimized instruction set used by the Beam to implement the external instruction set. To give you an understanding of how the Beam works we will cover this instruction set in [BEAM Internal Instructions](#). The internal instruction set can change without warning between minor version or even in patch releases. Basing any tool on the internal instruction set is risky.

In this chapter I will go through the general syntax for the instructions and some instruction groups in detail, a complete list of instructions with short descriptions can be found in [BEAM Instructions](#).

## 7.1. Instruction definitions

The names and opcodes of the generic instructions are defined in lib/compiler/src/genop.tab.

The file contains a version number for the Beam instruction format, which also is written to .beam files. This number has so far never changed and is still at version 0. If the external format would be changed in a non backwards compatible way this number would be changed.

The file genop.tab is used as input by beam\_makeops which is a perl script which generate code from the ops tabs. The generator is used both to generate Erlang code for the compiler (beam\_opcodes.hrl and beam\_opcodes.erl) and to generate C code for the emulator ( TODO: Filenames).

Any line in the file starting with "#" is a comment and ignored by beam\_makeops. The file can contain definitions, which turns into a binding in the perl script, of the form:

```
NAME=EXPR
```

Like, e.g.:

```
BEAM_FORMAT_NUMBER=0
```

The Beam format number is the same as the instructionset field in the external beam format. It is only bumped when a backwards incompatible change to the instruction set is made.

The main content of the file are the opcode definitions of the form:

```
OPNUM: [-]NAME/ARITY
```

Where OPNUM and ARITY are integers, NAME is an identifier starting with a lowercase letter (a-z), and :, -, and / are literals.

For example:

```
1: label/1
```

The minus sign (-) indicates a deprecated function. A deprecated function keeps its opcode in order for the loader to be sort of backwards compatible (it will recognize deprecated instructions and refuse to load the code).

In the rest of this Chapter we will go through some BEAM instructions in detail. For a full list with brief descriptions see: [BEAM Instructions](#).

## 7.2. BEAM code listings

As we saw in [The Compiler](#) we can give the option 'S' to the Erlang compiler to get a .S file with the BEAM code for the module in a human and machine readable format (actually as Erlang terms).

Given the file beamexample1.erl:

```
-module(beamexample1).  
-export([id/1]).  
  
id(I) when is_integer(I) -> I.
```

When compiled with erlc -S beamexample.erl we get the following beamexmaple.S file:

```

{module, beamexample1}. %% version = 0

{exports, [{id,1},{module_info,0},{module_info,1}]}.

{attributes, []}.

{labels, 7}.

{function, id, 1, 2}.
{label,1}.
{line,[{location,"beamexample1.erl",5}]}.
{func_info,{atom,beamexample1},{atom,id},1}.
{label,2}.
{test,is_integer,{f,1},[{x,0}]}.
return.

{function, module_info, 0, 4}.
{label,3}.
{line,[]}.
{func_info,{atom,beamexample1},{atom,module_info},0}.
{label,4}.
{move,{atom,beamexample1},{x,0}}.
{line,[]}.
{call_ext_only,1,{extfunc,erlang,get_module_info,1}}.

{function, module_info, 1, 6}.
{label,5}.
{line,[]}.
{func_info,{atom,beamexample1},{atom,module_info},1}.
{label,6}.
{move,{x,0},{x,1}}.
{move,{atom,beamexample1},{x,0}}.
{line,[]}.
{call_ext_only,2,{extfunc,erlang,get_module_info,2}}.

```

In addition to the actual beam code for the integer identity function we also get some meta instructions.

The first line {module, beamexample1}. %% version = 0 tells us the module name "beamexample1" and the version number for the instruction set "0".

Then we get a list of exported functions "id/1, module\_info/0, module\_info/1". As we can see the compiler has added two auto generated functions to the code. These two functions are just dispatchers to the generic module info BIFs (erlang:module\_info/1 and erlang:module\_info/2) with the name of the module added as the first argument.

The line {attributes, []} list all defined compiler attributes, none in our case.

Then we get to know that there are less than 7 labels in the module, {labels, 7}, which makes it easy to do code loading in one pass.

The last type of meta instruction is the function instruction on the format {function, Name, Arity, StartLabel}. As we can see with the id function the start label is actually the second label in the code of the function.

The instruction `{label, N}` is not really an instruction, it does not take up any space in memory when loaded. It is just to give a local name (or number) to a position in the code. Each label potentially marks the beginning of a basic block since it is a potential destination of a jump.

The first two instructions following the first label (`{label,1}`) are actually error generating code which adds the line number and module, function and arity information and throws an exception. That are the instructions `line` and `func_info`.

The meat of the function is after `{label,2}`, the instruction `{test,is_integer,{f,1},[{x,0}]}.` The test instruction tests if its arguments (in the list at the end, that is variable `{x,0}` in this case) fulfills the test, in this case is an integer (`is_integer`). If the test succeeds the next instruction (`return`) is executed. Otherwise the functions fails to label 1 (`{f,1}`), that is, execution continues at label one where a function clause exception is thrown.

The other two functions in the file are auto generated. If we look at the second function the instruction `{move,{x,0},{x,1}}` moves the argument in register `x0` to the second argument register `x1`. Then the instruction `{move,{atom,beamexample1},{x,0}}` moves the module name atom to the first argument register `x0`. Finally a tail call is made to `erlang:get_module_info/2` (`{call_ext_only,2,{extfunc,erlang,get_module_info,2}}`). As we will see in the next section there are several different call instructions.

## 7.3. Calls

As we have seen in [Different Types of Calls, Linking and Hot Code Loading \(5p\)](#) there are several different types of calls in Erlang. To distinguish between local and remote calls in the instruction set, remote calls have `_ext` in their instruction names. Local calls just have a label in the code of the module, while remote calls takes a destination of the form `{extfunc, Module, Function, Arity}`.

To distinguish between ordinary (stack building) calls and tail-recursive calls, the latter have either `_only` or `_last` in their name. The variant with `_last` will also deallocate as many stack slot as given by the last argument.

There is also a `call_fun Arity` instruction that calls the closure stored in register `{x, Arity}`. The arguments are stored in `x0` to `{x, Arity-1}`.

For a full listing of all types of call instructions see [BEAM Instructions](#).

## 7.4. Stack (and Heap) Management

The stack and the heap of an Erlang process on Beam share the same memory area see [Processes](#) and [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#) for a full discussion. The stack grows toward lower addresses and the heap toward higher addresses. Beam will do a garbage collection if more space than what is available is needed on either the stack or the heap.

### A leaf function

A leaf function is a function which doesn't call any other function.

### A non leaf function

A non leaf function is a function which may call another function.

On entry to a non leaf function the *continuation pointer* (CP) is saved on the stack, and on exit it is read back from the stack. This is done by the allocate and deallocate instructions, which are used for setting up and tearing down the stack frame for the current instruction.

A function skeleton for a leaf function looks like this:

```
{function, Name, Arity, StartLabel}.
{label,L1}.
{func_info,{atom,Module},{atom,Name},Arity}.
{label,L2}.
...
return.
```

A function skeleton for a non leaf function looks like this:

```
{function, Name, Arity, StartLabel}.
{label,L1}.
{func_info,{atom,Module},{atom,Name},Arity}.
{label,L2}.
{allocate,Need,Live}.

...
call ...
...

{deallocate,Need}.
return.
```

The instruction allocate StackNeed Live saves the continuation pointer (CP) and allocate space for StackNeed extra words on the stack. If a GC is needed during allocation save Live number of X registers. E.g. if Live is 2 then registers X0 and X1 are saved.

When allocating on the stack, the stack pointer (E) is decreased.

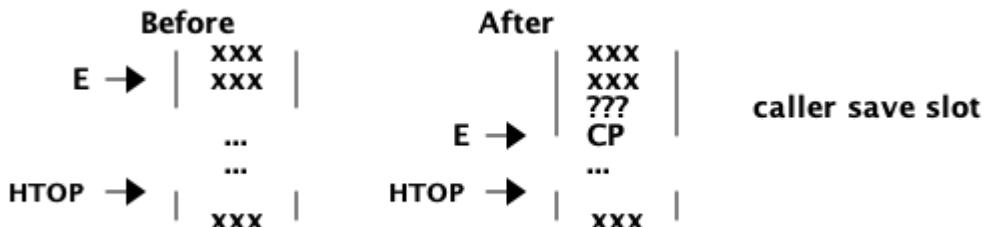


Figure 22. Allocate 10

For a full listing of all types of allocate and deallocate instructions see [BEAM Instructions](#).

## 7.5. Message Passing

Sending a message is straight forward in beam code. You just use the send instruction. Note though that the send instruction does not take any arguments, it is more like a function call. It assumes that the arguments (the destination and the message) are in the argument registers X0 and X1. The message is also copied from X1 to X0.

Receiving a message is a bit more complicated since it involves both selective receive with pattern matching and introduces a yield/resume point within a function body. (There is also a special feature to minimize message queue scanning using refs, more on that later.)

### 7.5.1. A Minimal Receive Loop

A minimal receive loop, which accepts any message and has no timeout (e.g. receive \_ -> ok end) looks like this in BEAM code:

```
{label,2}.
{wait,{f,1}}.
{label,1}.
{loop_rec,{f,2},{x,0}}.
remove_message.
{jump,{f,3}}.
{label,2}.
{wait,{f,1}}.
{label,3}.
...
...
```

The loop\_rec L2 x0 instruction first checks if there is any message in the message queue. If there are no messages execution jumps to L2, where the process will be suspended waiting for a message to arrive.

If there is a message in the message queue the loop\_rec instruction also moves the message from the *m-buf* to the process heap. See [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#) and [Processes](#) for details of the m-buf handling.

For code like receive \_ -> ok end, where we accept any messages, there is no pattern matching needed, we just do a remove\_message which unlinks the next message from the message queue. (It also removes any timeout, more on this soon.)

### 7.5.2. A Selective Receive Loop

For a selective receive like e.g. receive [] -> ok end we will loop over the message queue to check if any message in the queue matches.

```
{label,1}.
{loop_rec,{f,3},{x,0}}.
{test,is_nil,{f,2},[{x,0}]}.
remove_message.
{jump,{f,4}}.
{label,2}.
{loop_rec_end,{f,1}}.
{label,3}.
{wait,{f,1}}.
{label,4}.
...
...
```

In this case we do a pattern match for Nil after the loop\_rec instruction if there was a message in the mailbox. If the message doesn't match we end up at L3 where the instruction loop\_rec\_end advances the save pointer to the next message (*p->msg.save = &(\*p->msg.save)->next*) and jumps back to L2.

If there are no more messages in the message queue the process is suspended by the wait instruction at L4 with the save pointer pointing to the end of the message queue. When the processes is rescheduled it will only look at new messages in the message queue (after the save point).

### 7.5.3. A Receive Loop With a Timeout

If we add a timeout to our selective receive the wait instruction is replaced by a wait\_timeout instruction followed by a timeout instruction and the code following the timeout.

```

{label,1}.
{loop_rec,{f,3},{x,0}}.
{test,is_nil,{f,2},[{x,0}]}.
remove_message.
{jump,{f,4}}.
{label,2}.
{loop_rec_end,{f,1}}.
{label,3}.
{wait_timeout,{f,1},{integer,1000}}.
timeout.
{label,4}.
...

```

The wait\_timeout instruction sets up a timeout timer with the given time (1000 ms in our example) and it also saves the address of the next instruction (the timeout) in p->def\_arg\_reg[0] and then when the timer is set, p->i is set to point to def\_arg\_reg.

This means that if no matching message arrives while the process is suspended a timeout will be triggered after 1 second and execution for the process will continue at the timeout instruction.

Note that if a message that doesn't match arrives in the mailbox, the process is scheduled for execution and will run the pattern matching code in the receive loop, but the timeout will not be canceled. It is the remove\_message code which also removes any timeout timer.

The timeout instruction resets the save point of the mailbox to the first element in the queue, and clears the timeout flag (F\_TIMO) from the PCB.

#### 7.5.4. The Synchronous Call Trick (aka The Ref Trick)

We have now come to the last version of our receive loop, where we use the ref trick alluded to earlier to avoid a long message box scan.

A common pattern in Erlang code is to implement a type of "remote call" with send and a receive between two processes. This is for example used by gen\_server. This code is often hidden behind a library of ordinary function calls. E.g., you call the function counter:increment(Counter) and behind the scene this turns into something like Counter ! {self(), inc}, receive {Counter, Count} -> Count end.

This is usually a nice abstraction to encapsulate state in a process. There is a slight problem though when the mailbox of the calling process has many messages in it. In this case the receive will have to check each message in the mailbox to find out that no message except the last matches the return message.

This can quite often happen if you have a server that receives many messages and for each message does a number of such remote calls, if there is no back throttle in place the servers message queue will fill up.

To remedy this there is a hack in ERTS to recognize this pattern and avoid scanning the whole message queue for the return message.

The compiler recognizes code that uses a newly created reference (ref) in a receive (see [\[ref\\_trick\\_code\]](#)), and emits code to avoid the long inbox scan since the new ref can not already be in the inbox.

```

Ref = make_ref(),
Counter ! {self(), inc, Ref},
receive
  {Ref, Count} -> Count
end.

```

This gives us the following skeleton for a complete receive, see [\[ref\\_receive\]](#).

```
{recv_mark,{f,3}}.
{call_ext,0,{extfunc,erlang,make_ref,0}}.
...
send.
{recv_set,{f,3}}.
{label,3}.
{loop_rec,{f,5},{x,0}}.
{test,is_tuple,{f,4},[{x,0}]}.
...
{test,is_eq_exact,{f,4},[{x,1},{y,0}]}.
...
remove_message.
...
{jump,{f,6}}.
{label,4}.
{loop_rec_end,{f,3}}.
{label,5}.
{wait,{f,3}}.
{label,6}.
```

The recv\_mark instruction saves the current position (the end msg.last) in msg.saved\_last and the address of the label in msg.mark

The recv\_set instruction checks that msg.mark points to the next instruction and in that case moves the save point (msg.save) to the last message received before the creation of the ref (msg.saved\_last). If the mark is invalid ( i.e. not equal to msg.save) the instruction does nothing.

**Chapter 8****Different Types of Calls, Linking and Hot Code Loading****(5p)**

Local calls, remote calls, closure calls, tuple calls, p-mod calls. The code server. Linking. Hot code loading, purging. (Overlap with Chapter 4, have to see what goes where.) Higher order functions, Implementation of higher order functions. Higher order functions and hot code loading. Higher order functions in a distributed system.

## 8.1. Hot Code Loading

In Erlang there is a semantic difference between a local function call and a remote function call. A remote call, that is a call to a function in a named module, is guaranteed to go to the latest loaded version of that module. A local call, an unqualified call to a function within the same module, is guaranteed to go to the same version of the code as the caller.

A call to a local function can be turned into a remote call by specifying the module name at the call site. This is usually done with the `?MODULE` macro as in `?MODULE:foo()`. A remote call to a non local module can not be turned into a local call, i.e. there is no way to guarantee the version of the callee in the caller.

This is an important feature of Erlang which makes *hot code loading* or *hot upgrades* possible. Just make sure you have a remote call somewhere in your server loop and you can then load new code into the system while it is running; when execution reaches the remote call it will switch to executing the new code.

A common way of writing server loops is to have a local call for the main loop and a code upgrade handler which does a remote call and possibly a state upgrade:

```
loop(State) ->
    receive
        upgrade ->
            NewState = ?MODULE:code_upgrade(State),
            ?MODULE:loop(NewState);
        Msg ->
            NewState = handle_msg(Msg, State),
            loop(NewState)
    end.
```

With this construct, which is basically what `gen_server` uses, the programmer has control over when and how a code upgrade is done.

The hot code upgrade is one of the most important features of Erlang which makes it possible to write servers that operates 24/7 year out and year in. It is also one of the main reasons why Erlang is dynamically typed. It is very hard in a statically typed language to give type for the `code_upgrade` function. (It is also hard to give the type of the loop function). These types will change in the future as the type of State changes to handle new features.

For a language implementer concerned with performance, the hot code loading functionality is a burden though. Since each call to or from a remote module can change to new code in the future it is very hard to do whole program optimization across module boundaries. (Hard but not impossible, there are solutions but so far I have not seen one fully implemented).

## 8.2. Code Loading

```
<!--  
Shouldn't Code Loading come before Hot Code Loading? Or are the two topics not related in that way? -  
bmacdonald  
-->
```

In the Erlang Runtime System the code loading is handled by the code server. The code server will call the lower level BIFs in the erlang module for the actual loading. But the code server also determines the purging policy.

The runtime system can keep two versions of each module, a *current* version and an *old* version. All fully qualified (remote) calls go to the current version. Local calls in the old version and return addresses on the stack can still go to the old version.

If a third version of a module is loaded and there are still processes running (have pointers on the stack to) the code server will kill those processes and purge the old code. Then the current version will become old and the third version will be loaded as the current version.

## Chapter 9

# The BEAM Loader

## 9.1. Transforming from Generic to Specific instructions

The BEAM loader does not just take the external beam format and writes it to memory. It also does a number of transformations on the code and translates from the external (generic) format to the internal (specific) format.

The code for the loader can be found in `beam_load.c` (in `erts/emulator/beam`) but most of the logic for the translations are in the file `ops.tab` (in the same directory).

The first step of the loader is to parse beam file, basically the same work as we did in Erlang in [Modules and The BEAM File Format \(16p\)](#) but written in C.

Then the rules in `ops.tab` are applied to instructions in the code chunk to translate the generic instruction to one or more specific instructions.

The translation table works through pattern matching. Each line in the file defines a pattern of one or more generic instructions with arguments and optionally an arrow followed by one or more instructions to translate to.

The transformations in `ops.tab` tries to handle patterns of instructions generated by the compiler and peephole optimize them to fewer specific instructions. The `ops.tab` transformations tries to generate jump tables for patterns of selects.

The file `ops.tab` is not parsed at runtime, instead a pattern matching program is generated from `ops.tab` and stored in an array in a generated C file. The perl script `beam_makeops` (in `erts/emulator/utils`) generates a target specific set of opcodes and translation programs in the files `beam_opcodes.h` and `beam_opcodes.c` (these files end up in the given target directory e.g. `erts/emulator/x86_64-unknown-linux-gnu/opt/smp/`).

The same program (`beam_makeops`) also generates the Erlang code for the compiler back end `beam_opcodes.erl`.

## 9.2. Understanding ops.tab

The transformations in `ops.tab` [<https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/ops.tab>] are executed in the order that they are written in the file. So just like in Erlang pattern matching, the different rules are triggered from top to bottom.

The types that `ops.tab` uses for arguments in instructions can be found in [BEAM Instructions](#).

### 9.2.1. Transformations

Most of the rules in `ops.tab` are transformations between different instructions. A simple transformation looks like this:

```
move S x==0 | return => move_return S
```

This combines a `move` from any location to `x(0)` and return into a single instruction called `move_return`. Let's break the transformation apart to see what the different parts do.

#### `move`

is the instruction that the pattern first has to match. This can be either a generic instruction that the compiler has emitted, or a temporary instruction that `ops.tab` has emitted to help with transformations.

**S**

is a variable binding any type of value. Any value in the pattern (left hand side or  $\Rightarrow$ ) that is used in the generator (right hand side of  $\Rightarrow$ ) has to be bound to a variable.

**x==0**

is a guard that says that we only apply the transformation if the target location is an **x** register with the value **0**. It is possible to chain multiple types and also bind a variable here. For instance **D=xy==0** would allow both **x** and **y** registers with a value of **0** and also bind the argument to the variable **D**.

|

signifies the end of this instruction and the beginning of another instruction that is part of the same pattern.

**return**

is the second instruction to match in this pattern.

 $\Rightarrow$ 

signifies the end of the pattern and the start of the code that is to be generated.

**move\_return S**

is the name of the generated instruction together with the name of the variable on the lhs. It is possible to generate **multiple instructions** [<https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/ops.tab#L625>] as part of a transformation by using the | symbol.

*A more complex example*

More complex translations can be done in **ops.tab**. For instance take the **select\_val** [<https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/ops.tab#L127-L182>] instruction. It will be translated by the loader into either a jump table, a linear search array or a binary search array depending on the input values.

```
is_integer Fail=f S | select_val S=s Fail=f Size=u Rest=* | \
use_jump_tab(Size, Rest) => gen_jump_tab(S, Fail, Size, Rest)
```

The above transformation creates a jump table if possible of the **select\_val**. There are a bunch of new techniques used in the transformations.

**S**

is used in both **is\_integer** and **select\_val**. This means that both the values have to be of the same type and have the same value. Furthermore the **S=s** guard limits the type to a be a source register.

**Rest=\***

allows a variable number of arguments in the instruction and binds them to variable **Rest**.

**use\_jump\_tab(Size, Rest)**

calls the **use\_jump\_tab** [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\\_load.c#L2707](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam_load.c#L2707)] C function in **beam\_load.c** that decides whether the arguments in the **select\_val** can be transformed into a jump table.

\

signifies that the transformation rule continues on the next line.

**gen\_jump\_tab(S, Fail, Size, Rest)**

calls the **gen\_jump\_tab** [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\\_load.c#L3692](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam_load.c#L3692)] C function in **beam\_load.c** that takes care of generating the appropriate instruction.

## 9.2.2. Specific instruction

When all transformations are done, we have to decide how the specific instruction should look like. Let's continue to look at `%macro: move_return`:

```
%macro: move_return MoveReturn -nonext
move_return x
move_return c
move_return n
```

This will generate three different instructions that will use the `MoveReturn` [https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\_emu.c#L636] macro in `beam_emu.c` to do the work.

### `%macro: move_return`

this tells `ops.tab` to generate the code for `move_return`. If there is no `%macro` line, the instruction has to be implemented by hand in `beam_emu.c`. The code for the instruction will be placed in `beam_hot.h` or `beam_cold.h` depending on if the `%hot` or `%cold` directive is active.

### `MoveReturn`

tells the code generator to that the name of the c-macro in `beam_emu.c` to use is `MoveReturn`. This macro has to be implemented manually.

### `-nonext`

tells the code generator that it should not generate a dispatch to the next instruction, the `MoveReturn` macro will take care of that.

### `move_return x`

tells the code generator to generate a specific instruction for when the instruction argument is an `x` register. `c` for when it is a constant, `n` when it is `NIL`. No instructions are generated for when the argument is a `y` register as the compiler will never generate such code.

The resulting code in `beam_hot.h` will look like this:

```
OpCase(move_return_c):
{
    MoveReturn(Arg(0));
}

OpCase(move_return_n):
{
    MoveReturn(NIL);
}

OpCase(move_return_x):
{
    MoveReturn(xb(Arg(0)));
}
```

All the implementor has to do is to define the `MoveReturn` macro in `beam_emu.c` and the instruction is complete.

### *Macro flags*

The `%macro` rules can take multiple different flags to modify the code that gets generated.

The examples below assume that there is a specific instruction looking like this:

```
%macro move_call MoveCall  
move_call x f
```

without any flags to the **%macro** we the following code will be generated:

```
BeamInstr* next;  
PreFetch(2, next);  
MoveCall(Arg(0));  
NextPF(2, next);
```



The **PreFetch** and **NextPF** [https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\_emu.c#L519-L523] macros make sure to load the address to jump to next before the instruction is executed. This trick increases performance on all architectures by a varying amount depending on cache architecture and super scalar properties of the CPU.

#### -nonext

Don't emit a dispatch for this instructions. This is used for instructions that are known to not continue with the next instructions, i.e. return, call, jump.

```
%macro move_call MoveCall -nonext
```

```
MoveCall(xb(Arg(0)));
```

#### -arg\_\*

Include the arguments of type \* as arguments to the c-macro. Not all argument types are included by default in the c-macro. For instance the type **f** used for fail labels and local function calls is not included. So giving the option **-arg\_f** will include that as an argument to the c-macro.

```
%macro move_call MoveCall -arg_f
```

```
MoveCall(xb(Arg(0)), Arg(1));
```

#### -size

Include the size of the instruction as an argument to the c-macro.

```
%macro move_call MoveCall -size
```

```
MoveCall(xb(Arg(0)), 2);
```

#### -pack

Pack any arguments if possible. This places multiple register arguments in the same word if possible. As register arguments can only be 0-1024, we only need 10 bits to store them + 2 for tagging. So on a 32-bit system we can put 2 registers in one word, while on a 64-bit we can put 4 registers in one word. Packing instruction can greatly decrease the memory used for a single instruction. However there is also a small cost to unpack the instruction, which is why it is not enabled for all instructions.

The example with the call cannot do any packing as **f** cannot be packed and only one other argument exists. So let's look at the **put\_list** [https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/ops.tab#L539] instruction as an example instead.

```
%macro:put_list PutList -pack
put_list x x x
```

```
BeamInstr tmp_packed1;
BeamInstr* next;
PreFetch(1, next);
tmp_packed1 = Arg(0);
PutList(xb(tmp_packed1&BEAM_TIGHT_MASK),
        xb((tmp_packed1>>BEAM_TIGHT_SHIFT)&BEAM_TIGHT_MASK),
        xb((tmp_packed1>>(2*BEAM_TIGHT_SHIFT))));
```

NextPF(1, next);

This packs the 3 arguments into 1 machine word, which halves the required memory for this instruction.

#### **-fail\_action**

Include a fail action as an argument to the c-macro. Note that the [ClauseFail\(\)](#) [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\\_emu.c#L2996-L2998](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam_emu.c#L2996-L2998)] macro assumes the fail label is in the first argument of the instructions, so in order to use this in the above example we should transform the `move_call x f` to `move_call f x`.

```
%macro move_call MoveCall -fail_action
```

```
MoveCall(xb(Arg(0)), ClauseFail());
```

#### **-gen\_dest**

Include a [store function](#) [[https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam\\_emu.c#L166-L174](https://github.com/erlang/otp/blob/OTP-19.3/erts/emulator/beam/beam_emu.c#L166-L174)] as an argument to the c-macro.

```
%macro move_call MoveCall -gen_dest
```

```
MoveCall(xb(Arg(0)), StoreSimpleDest);
```

#### **-goto**

Replace the normal next dispatch with a jump to a c-label inside beam\_emu.c

```
%macro move_call MoveCall -goto:do_call
```

```
MoveCall(xb(Arg(0)));
goto do_call;
```

## 9.3. Optimizations

The loader performs many peephole optimizations when loading the code. The most important ones are instruction combining and instruction specialization.

Instruction combining is the joining of two or more smaller instructions into one larger instruction. This can lead to a large speed up of the code if the instructions are known to follow each other most of the time. The speed up is achieved because there is no longer any need to do a dispatch inbetween the instructions, and also the C compiler gets more information to work with when it is optimizing that instruction. When to do instruction combining is a trade-off where one has to consider the impact the increased size

of the main emulator loop has vs the gain when the instruction is executed.

Instruction specialization removes the need to decode the arguments in an instruction. So instead of having one `move_sd` instruction, `move_xx`, `move_xy` etc are generated with the arguments already decoded. This reduces the decode cost of the instructions, but again it is a tradeoff vs emulator code size.

### 9.3.1. select\_val optimizations

The `select_val` instruction is emitted by the compiler to do control flow handling of many functions or case clauses. For instance:

```
select(1) -> 3;
select(2) -> 3;
select(_) -> error.
```

compiles to:

```
{function, select, 1, 2}.
{label,1}.
{line,[{location,"select.erl",5}]}.
{func_info,{atom,select},{atom,select},1}.
{label,2}.
{test,is_integer,{f,4},[{x,0}]}.
{select_val,{x,0},{f,4},{list,[{integer,2},{f,3},{integer,1},{f,3}]}}.
{label,3}.
{move,{integer,3},{x,0}}.
return.
{label,4}.
{move,{atom,error},{x,0}}.
return.
```

The values in the condition are only allowed to be either integers or atoms. If the value is of any other type the compiler will not emit a `select_val` instruction. The loader uses a couple of heuristics to figure out what type algorithm to use when doing the `select_val`.

#### jump\_on\_val

Create a jump table and use the value as the index. This is very efficient and happens when a group of close together integers are used as the value to select on. If not all values are present, the jump table is padded with extra fail label slots.

#### select\_val2

Used when only two values are to be selected upon and they do not fit in a jump table.

#### select\_val\_lins

Do a linear search of the sorted atoms or integers. This is used when a small amount of atoms or integers are to be selected from.

#### select\_val\_bins

Do a binary search of the sorted atoms or integers.

### 9.3.2. pre-hashing of literals

When a literal is loaded and used as an argument to any of the bifs or instructions that need a hashed value of the literal, instead of hashing the literal value every time, the hash is created by the loader and used by the instructions.

Examples of code using this technique is maps instructions and also the process dictionary bifs.



## Chapter 10

---

# BEAM Internal Instructions



# Chapter 11

# Scheduling

To fully understand where time in an ERTS system is spent you need to understand how the system decides which Erlang code to run and when to run it. These decisions are made by the Scheduler.

The scheduler is responsible for the real-time guarantees of the system. In a strict Computer Science definition of the word real-time, a real-time system has to be able to guarantee a response within a specified time. That is, there are real deadlines and each task has to complete before its deadline. In Erlang there are no such guarantees, a *timeout* in Erlang is only guaranteed to **not** trigger **before** the given deadline.

In a general system like Erlang where we want to be able to handle all sorts of programs and loads, the scheduler will have to make some compromises. There will always be corner cases where a generic scheduler will behave badly. After reading this chapter you will have a deeper understanding of how the Erlang scheduler works and especially when it might not work optimally. You should be able to design your system to avoid the corner cases and you should also be able to analyze a misbehaving system.

## 11.1. Concurrency, Parallelism, and Preemptive Multitasking

Erlang is a concurrent language. When we say that processes run concurrently we mean that for an outside observer it looks like two processes are executing at the same time. In a single core system this is achieved by preemptive multitasking. This means that one process will run for a while, and then the scheduler of the virtual machine will suspend it and let another process run.

In a multicore or a distributed system we can achieve true parallelism, that is, two or more processes actually executing at the exact same time. In an SMP enabled emulator the system uses several OS threads to indirectly execute Erlang processes by running one scheduler and emulator per thread. In a system using the default settings for ERTS there will be one thread per enabled core (physical or hyper threaded).

We can check that we have a system capable of parallel execution, by checking if SMP support is enabled:

```
iex(1)> :erlang.system_info :smp_support
true
```

We can also check how many schedulers we have running in the system:

```
iex(2)> :erlang.system_info :.schedulers_online
4
```

We can see this information in the Observer as shown in the figure below.

If we spawn more processes than schedulers we have and let them do some busy work we can see that there are a number of processes *running* in parallel and some processes that are *runnable* but not currently running. We can see this with the function `erlang:process_info/2`.

```

1> Loop = fun (0, _) -> ok; (N, F) -> F(N-1, F) end,
BusyFun = fun() -> spawn(fun () -> Loop(1000000, Loop) end) end,
SpawnThem = fun(N) -> [ BusyFun() || _ <- lists:seq(1, N)] end,
GetStatus = fun() -> lists:sort([{erlang:process_info(P, [status]), P}
                                  || P <- erlang:processes()]) end,
RunThem = fun (N) -> SpawnThem(N), GetStatus() end,
RunThem(8).

[{{status,garbage_collecting}],<0.62.0>},
[{{status,garbage_collecting}],<0.66.0>},
[{{status,runnable}],<0.60.0>},
[{{status,runnable}],<0.61.0>},
[{{status,runnable}],<0.63.0>},
[{{status,runnable}],<0.65.0>},
[{{status,runnable}],<0.67.0>},
[{{status,running}],<0.58.0>},
[{{status,running}],<0.64.0>},
[{{status,waiting}],<0.0.0>},
[{{status,waiting}],<0.1.0>],
...

```

We will look closer at the different statuses that a process can have later in this chapter, but for now all we need to know is that a process that is *running* or *garbage\_collecting* is actually running in on a scheduler. Since the machine in the example has four cores and four schedulers there are four process running in parallel (the shell process and three of the *busy processes*). There are also five busy processes waiting to run in the state *runnable*.

By using the *Load Charts* tab in the Observer we can see that all four schedulers are fully loaded while the busy processes execute.

```

observer:start().
ok
3> RunThem(8).

```

[Observer] | ..../images/observer\_load.jpg

## 11.2. Preemptive Multitasking in ERTS Cooperating in C

The preemptive multitasking on the Erlang level is achieved by cooperative multitasking on the C level. The Erlang language, the compiler and the virtual machine works together to ensure that the execution of an Erlang process will yield within a limited time and let the next process run. The technique used to measure and limit the allowed execution time is called reduction counting, we will look at all the details of reduction counting soon.

## 11.3. Reductions

One can describe the scheduling in BEAM as preemptive scheduling on top of cooperative scheduling. A process can only be suspended at certain points of the execution, such as at a receive or a function call. In that way the scheduling is cooperative---a process has to execute code which allows for suspension. The nature of Erlang code makes it almost impossible for a process to run for a long time without doing a function call. There are a few Built In Functions (BIFs) that still can take too long without yielding. Also, if you call C code in a badly implemented Native Implemented Function (NIF) you might block one scheduler for a long time. We will look at how to write well behaved NIFs in [Interfacing C — BIFs NIFs and Linked in Drivers](#).

Since there are no other loop constructs than recursion and list comprehensions, there is no way to loop forever without doing a

function call. Each function call is counted as a **reduction**; when the reduction limit for the process is reached it is suspended.



#### *Version Info*

Prior to OTP-20.0, the value of **CONTEXT\_REDs** was 2000.



#### *Reductions*

The term reduction comes from the Prolog ancestry of Erlang. In Prolog each execution step is a goal-reduction, where each step reduces a logic problem into its constituent parts, and then tries to solve each part.

### 11.3.1. How Many Reductions Will You Get?

When a process is scheduled it will get a number of reductions defined by **CONTEXT\_REDs** (defined in `erl_vm.h` [[https://github.com/erlang/otp/blob/OTP-20.0/erts/emulator/beam/erl\\_vm.h](https://github.com/erlang/otp/blob/OTP-20.0/erts/emulator/beam/erl_vm.h)], currently as 4000). After using up its reductions or when doing a receive without a matching message in the inbox, the process will be suspended and a new processes will be scheduled.

If the VM has executed as many reductions as defined by **INPUT\_REDUCTIONS** (currently  $2 * \text{CONTEXT\_REDs}$ , also defined in `erl_vm.h`) or if there is no process ready to run the scheduler will do system-level activities. That is, basically, check for IO; we will cover the details soon.

### 11.3.2. What is a Reduction Really?

It is not completely defined what a reduction is, but at least each function call should be counted as a reduction. Things get a bit more complicated when talking about BIFs and NIFs. A process should not be able to run for "a long time" without using a reduction and yielding. A function written in C can not yield in the middle, it has to make sure it is in a clean state and return. In order to be re-entrant it has to save its internal state somehow before it returns and then set up the state again on re-entry. This can be very costly, especially for a function that sometimes only does little work and sometimes lot. The reason for writing a function in C instead of Erlang is usually to achieve performance and to not do unnecessary book keeping work. Since there is no clear definition of what one reduction is, other than a function call on the Erlang level, there is a risk that a function implemented in C takes many more clock cycles per reduction than a normal Erlang function. This can lead to an imbalance in the scheduler, and even starvation.

For example in Erlang versions prior to R16, the BIFs `binary_to_term/1` and `term_to_binary/1` were non yielding and only counted as one reduction. This meant that a process calling these functions on large terms could starve other processes. This can even happen in a SMP system because of the way processes are balanced between schedulers, which we will get to soon.

While a process is running the emulator keeps the number of reductions left to execute in the (register mapped) variable **FCALLS** (see `beam_emu.c`).

We can examine this value with `hipe_bifs:show_pcb/1`:

```
iex(13)> :hipe_bifs.show_pcb self
P: 0x00007efd7c2c0400
-----
Offset| Name          | Value      | *Value |
  0   | id            | 0x00000270000004e3 |         |
...
  328 | rcount        | 0x0000000000000000 |         |
  336 | reds          | 0x000000000000a528 |         |
...
  320 | fcalls        | 0x00000000000004a3 |         |
```

The field `reds` keep track of the total number of reductions a process has done up until it was last suspended. By monitoring this number you can see which processes do the most work.

You can see the total number of reductions for a process (the `reds` field) by calling `erlang:process_info/2` with the atom `reductions` as the second argument. You can also see this number in the process tab in the observer or with the i/o command in the Erlang shell.

As noted earlier, each time a process starts the field `fcalls` is set to the value of `CONTEXT_RED` and for each function call the process executes `fcalls` is reduced by 1. When the process is suspended the field `reds` is increased by the number of executed reductions. In some C like code something like: `p->reds += (CONTEXT_RED - p->fcalls)`.

Normally a process would do all its allotted reductions and `fcalls` would be 0 at this point, but if the process suspends in a receive waiting for a message it will have some reductions left.

When a process uses up all its reductions it will yield to let another process run, it will go from the process state *running* to the state *runnable*, if it yields in a receive it will instead go into the state *waiting* (for a message). In the next section we will take a look at all the different states a process can be in.

## 11.4. The Process State (or *status*)

The field `status` in the PCB contains the process state. It can be one of *free*, *runnable*, *waiting*, *running*, *exiting*, *garbing*, and *suspended*. When a process exits it is marked as free---you should never be able to see a process in this state, it is a short lived state where the process no longer exist as far as the rest of the system is concerned but there is still some clean up to be done (freeing memory and other resources).

Each process status represents a state in the Process State Machine. Events such as a timeout or a delivered message triggers transitions along the edges in the state machine. The *Process State Machine* looks like this:

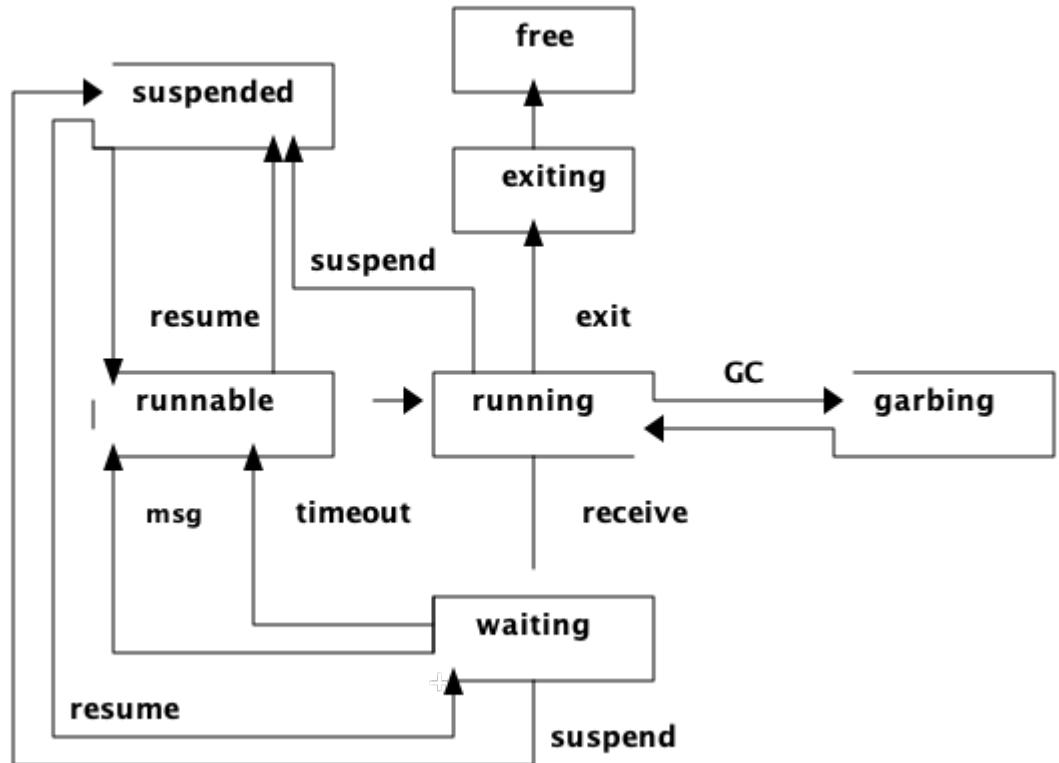


Figure 23. Process State Machine

The normal states for a process are *runnable*, *waiting*, and *running*. A running process is currently executing code in one of the schedulers. When a process enters a receive and there is no matching message in the message queue, the process will become waiting until a message arrives or a timeout occurs. If a process uses up all its reductions, it will become runnable and wait for a scheduler to pick it up again. A waiting process receiving a message or a timeout will become runnable.

Whenever a process needs to do garbage collection, it will go into the *garbing* state until the GC is done. While it is doing GC it saves the old state in the field `gcstatus` and when it is done it sets the state back to the old state using `gcstatus`.

The suspended state is only supposed to be used for debugging purposes. You can call `erlang:suspend_process/2` on another process to force it into the suspended state. Each time a process calls `suspend_process` on another process, the *suspend count* is increased. This is recorded in the field `rcount`. A call to (`erlang:resume_process/1`) by the suspending process will decrease the suspend count. A process in the suspend state will not leave the suspend state until the suspend count reaches zero.

The field `rstatus` (resume status) is used to keep track of the state the process was in before a suspend. If it was *running* or *runnable* it will start up as *runnable*, and if it was *waiting* it will go back to the wait queue. If a suspended waiting process receives a timeout `rstatus` is set to *runnable* so it will resume as *runnable*.

To keep track of which process to run next the scheduler keeps the processes in a queue.

## 11.5. Process Queues

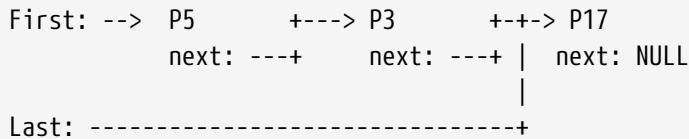
The main job of the scheduler is to keep track of work queues, that is, queues of processes and ports.

There are two process states that the scheduler has to handle, *runnable*, and *waiting*. Processes waiting to receive a message are in the waiting state. When a waiting process receives a message the send operations triggers a move of the receiving process into the *runnable* state. If the receive statement has a timeout the scheduler has to trigger the state transition to *runnable* when the timeout triggers. We will cover this mechanism later in this chapter.

### 11.5.1. The Ready Queue

Processes in the runnable state are placed in a FIFO (first in first out) queue handled by the scheduler, called the *ready queue*. The queue is implemented by a first and a last pointer and by the next pointer in the PCB of each participating process. When a new process is added to the queue the *last* pointer is followed and the process is added to the end of the queue in an O(1) operation. When a new process is scheduled it is just popped from the head (the *first* pointer) of the queue.

#### The Ready Queue



In a SMP system, where you have several scheduler threads, there is one queue per scheduler.

Scheduler 1	Scheduler 2	Scheduler 3	Scheduler 4
Ready: P5 P3 P17	Ready: P1 P4	Ready: P7 P12	Ready: P9 P10

The reality is slightly more complicated since Erlang processes have priorities. Each scheduler actually has three queues. One queue for *max priority* tasks, one for *high priority* tasks and one queue containing both *normal* and *low priority* tasks.

Scheduler 1	Scheduler 2	Scheduler 3	Scheduler 4
Max: P5 High: Normal: P3 P17	Max: High: P1 Ready: P4	Max: High: Ready: P7 P12 P10	Max: High: Ready: P9

If there are any processes in the max queue the scheduler will pick these processes for execution. If there are no processes in the max queue but there are processes in the high priority queue the scheduler will pick those processes. Only if there are no processes in the max and the high priority queues will the scheduler pick the first process from the normal and low queue.

When a normal process is inserted into the queue it gets a *schedule count* of 1 and a low priority process gets a schedule count of 8. When a process is picked from the front of the queue its schedule count is reduced by one, if the count reaches zero the process is scheduled, otherwise it is inserted at the end of the queue. This means that low priority processes will go through the queue seven times before they are scheduled.

### 11.5.2. Waiting, Timeouts and the Timing Wheel

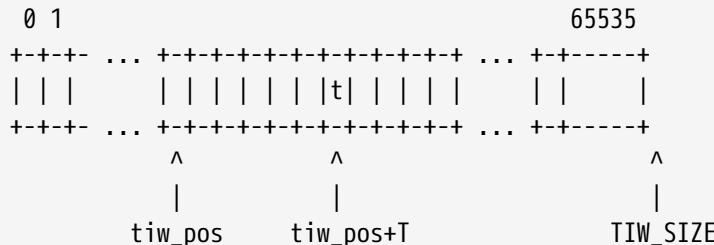
A process trying to do a receive on an empty mailbox or on a mailbox with no matching messages will yield and go into the waiting state.

When a message is delivered to an inbox the sending process will check whether the receiver is *sleeping* in the waiting state, and in that case it will *wake* the process, change its state to runnable, and put it at the end of the appropriate ready queue.

If the receive statement has a timeout clause a timer will be created for the process which will trigger after the specified timeout time. The only guarantee the runtime system gives on a timeout is that it will not trigger before the set time, it might be some time after the intended time before the process is scheduled and gets to execute.

Timers are handled in the VM by a *timing wheel*. That is, an array of time slots which wraps around. Prior to Erlang 18 the timing wheel was a global resource and there could be some contention for the write lock if you had many processes inserting timers into the wheel. Make sure you are using a later version of Erlang if you use many timers.

The default size (TIW\_SIZE) of the timing wheel is 65536 slots (or 8192 slots if you have built the system for a small memory footprint). The current time is indicated by an index into the array (tiw\_pos). When a timer is inserted into the wheel with a timeout of T the timer is inserted into the slot at (tiw\_pos+T)%TIW\_SIZE.



The timer stored in the timing wheel is a pointer to an ErlTimer struct. See [erl\\_time.h](https://github.com/erlang/otp/blob/OTP-19.1/erts/emulator/beam/erl_time.h) [https://github.com/erlang/otp/blob/OTP-19.1/erts/emulator/beam/erl\_time.h]. If several timers are inserted into the same slot they are linked together in a linked list by the prev and next fields. The count field is set to T/TIW\_SIZE

```

/*
** Timer entry:
*/
typedef struct erl_timer {
    struct erl_timer* next; /* next entry tiw slot or chain */
    struct erl_timer* prev; /* prev entry tiw slot or chain */
    Uint slot;             /* slot in timer wheel */
    Uint count;            /* number of loops remaining */
    int active;            /* 1=activated, 0=deactivated */
    /* called when timeout */
    void (*timeout)(void*);
    /* called when cancel (may be NULL) */
    void (*cancel)(void*);
    void* arg;              /* argument to timeout/cancel procs */
} ErlTimer;

```

## 11.6. Ports

A port is an Erlang abstraction for a communication point with the world outside of the Erlang VM. Communications with sockets, pipes, and file IO are all done through ports on the Erlang side.

A port, like a process, is created on the same scheduler as the creating process. Also like processes ports use reductions to decide when to yield, and they also get to run for 4000 reductions. But since ports don't run Erlang code there are no Erlang function calls to count as reductions, instead each *port task* is counted as a number of reductions. Currently a task uses a little more than 200 reductions per task, and a number of reductions relative to one thousands of the size of transmitted data.

A port task is one operation on a port, like opening, closing, sending a number of bytes or receiving data. In order to execute a port task the executing thread takes a lock on the port.

Port tasks are scheduled and executed in each iteration in the scheduler loop (see below) before a new process is selected for execution.

## 11.7. Reductions

When a process is scheduled it will get a number of reductions defined by `CONTEXT_REDs` (defined in `erl_vm.h` [[https://github.com/erlang/otp/blob/OTP-20.0/erts/emulator/beam/erl\\_vm.h](https://github.com/erlang/otp/blob/OTP-20.0/erts/emulator/beam/erl_vm.h)], currently as 4000). After using up its reductions or when doing a receive without a matching message in the inbox, the process will be suspended and a new processes will be scheduled.

If the VM has executed as many reductions as defined by `INPUT_REDUs` (currently  $2 * \text{CONTEXT\_REDs}$ , also defined in `erl_vm.h`) or if there is no process ready to run the scheduler will do system-level activities. That is, basically, check for IO; we will cover the details soon.

It is not completely defined what a reduction is, but at least each function call should be counted as a reduction. Things get a bit more complicated when talking about BIFs and NIFs. A process should not be able to run for "a long time" without using a reduction and yielding. A function written in C can usually not yield at any time, and the reason for writing it in C is usually to achieve performance. In such functions a reduction might take longer which can lead to imbalance in the scheduler.

For example in Erlang versions prior to R16 the BIFs `binary_to_term/1` and `term_to_binary/1` where non yielding and only counted as one reduction. This meant that a process calling these functions on large terms could starve other processes. This can even happen in a SMP system because of the way processes are balanced between schedulers, which we will get to soon.

While a process is running the emulator keeps the number of reductions left to execute in the (register mapped) variable `FCALLS` (see `beam_emu.c`).

## 11.8. The Scheduler Loop

Conceptually you can look at the scheduler as the driver of program execution in the Erlang VM. In reality, that is, the way the C code is structured, it is the emulator (`process_main` in `beam_emu.c`) that drives the execution and it calls the scheduler as a subroutine to find the next process to execute.

Still, we will pretend that it is the other way around, since it makes a nice conceptual model for the scheduler loop. That is, we see it as the scheduler picking a process to execute and then handing over the execution to the emulator.

Looking at it that way, the scheduler loop looks like this:

1. Update reduction counters.
2. Check timers
3. If needed check balance
4. If needed migrate processes and ports
5. Do auxiliary scheduler work
6. If needed check IO and update time
7. While needed pick a port task to execute
8. Pick a process to execute

## 11.9. Load Balancing

The current strategy of the load balancer is to use as few schedulers as possible without overloading any CPU. The idea is that you will get better performance through better memory locality when processes share the same CPU.

One thing to note though is that the load balancing done in the scheduler is between scheduler threads and not necessarily between CPUs or cores. When you start the runtime system you can specify how schedulers should be allocated to cores. The default behaviour is that it is up to the OS to allocate scheduler threads to cores, but you can also choose to bind schedulers to cores.

The load balancer assumes that there is one scheduler running on each core so that moving a process from a overloaded scheduler to an under utilized scheduler will give you more parallel processing power. If you have changed how schedulers are allocated to

cores, or if your OS is overloaded or bad at assigning threads to cores, the load balancing might actually work against you.

The load balancer uses two techniques to balance the load, *task stealing* and *migration*. Task stealing is used every time a scheduler runs out of work, this technique will result in the work becoming more spread out between schedulers. Migration is more complicated and tries to compact the load to the right number of schedulers.

### 11.9.1. Task Stealing

If a scheduler run queue is empty when it should pick a new process to schedule the scheduler will try to steal work from another scheduler.

First the scheduler takes a lock on itself to prevent other schedulers to try to steal work from the current scheduler. Then it checks if there are any inactive schedulers that it can steal a task from. If there are no inactive schedulers with stealable tasks then it will look at active schedulers, starting with schedulers having a higher id than itself, trying to find a stealable task.

The task stealing will look at one scheduler at a time and try to steal the highest priority task of that scheduler. Since this is done per scheduler there might actually be higher priority tasks that are stealable on another scheduler which will not be taken.

The task stealing tries to move tasks towards schedulers with lower numbers by trying to steal from schedulers with higher numbers, but since the stealing also will wrap around and steal from schedulers with lower numbers the result is that processes are spread out on all active schedulers.

Task stealing is quite fast and can be done on every iteration of the scheduler loop when a scheduler has run out of tasks.

### 11.9.2. Migration

To really utilize the schedulers optimally a more elaborate migration strategy is used. The current strategy is to compact the load to as few schedulers as possible, while at the same time spread it out so that no scheduler is overloaded.

This is done by the function *check\_balance* in *erl\_process.c*.

The migration is done by first setting up a migration plan and then letting schedulers execute on that plan until a new plan is set up. Every 2000\*CONTEXT\_REDOS reductions a scheduler calculates a migration path per priority per scheduler by looking at the workload of all schedulers. The migration path can have three different types of values: 1) cleared 2) migrate to scheduler # 3) immigrate from scheduler #

When a process becomes ready (for example by receiving a message or triggering a timeout) it will normally be scheduled on the last scheduler it ran on (S1). That is, if the migration path of that scheduler (S1), at that priority, is cleared. If the migration path of the scheduler is set to emigrate (to S2) the process will be handed over to that scheduler if both S1 and S2 have unbalanced run-queues. We will get back to what that means.

When a scheduler (S1) is to pick a new process to execute it checks to see if it has an immigration path from (S2) set. If the two involved schedulers have unbalanced run-queues S1 will steal a process from S2.

The migration path is calculated by comparing the maximum run-queues for each scheduler for a certain priority. Each scheduler will update a counter in each iteration of its scheduler loop keeping track of the maximal queue length. This information is then used to calculate an average (max) queue length (*AMQL*).

```

Max
Run Q
Length
 5      0
      0
      0  0
Avg: 2.5 -----
      0  0    0
 1      0  0    0

scheduler S1 S2 S3 S4

```

Then the schedulers are sorted on their max queue lengths.

```

Max
Run Q
Length
 5      0
      0
      0  0
Avg: 2.5 -----
      0  0    0
 1      0  0    0

scheduler S3 S4 S1 S2
      ^      ^
      |      |
    tix     fix

```

Any scheduler with a longer run queue than average (S1, S2) will be marked for emigration and any scheduler with a shorter max run queue than average (S3, S4) will be targeted for immigration.

This is done by looping over the ordered set of schedulers with two indices (immigrate from (fix)) and (emigrate to (tix)). In each iteration of the a loop the immigration path of S[tix] is set to S[fix] and the emigration path of S[fix] is set to S[tix]. Then tix is increased and fix decreased till they both pass the balance point. If one index reaches the balance point first it wraps.

In the example:

- \* Iteration 1: S2.emigrate\_to = S3 and S3.immigrate\_from = S2
- \* Iteration 2: S1.emigrate\_to = S4 and S4.immigrate\_from = S1

Then we are done.

In reality things are a bit more complicated since schedulers can be taken offline. The migration planning is only done for online schedulers. Also, as mentioned before, this is done per priority level.

When a process is to be inserted into a ready queue and there is a migration path set from S1 to S2 the scheduler first checks that the run queue of S1 is larger than AMQL and that the run queue of S2 is smaller than the average. This way the migration is only allowed if both queues are still unbalanced.

There are two exceptions though where a migration is forced even when the queues are balanced or even imbalanced in the wrong way. In both these cases a special evacuation flag is set which overrides the balance test.

The evacuation flag is set when a scheduler is taken offline to ensure that no new processes are scheduled on an offline scheduler. The flag is also set when the scheduler detects that no progress is made on some priority. That is, if there for example is a max

priority process which always is ready to run so that no normal priority processes ever are scheduled. Then the evacuation flag will be set for the normal priority queue for that scheduler.



**Chapter 12**

# The Memory Subsystem: Stacks, Heaps and Garbage Collection

Before we dive into the memory subsystem of ERTS, we need to have some basic vocabulary and understanding of the general memory layout of a program in a modern operating system. In this review section I will assume the program is compiled to an ELF executable and running on Linux on something like an IA-32/AMD64 architecture. The layout and terminology is basically the same for all operating systems that ERTS compile on.

A program's memory layout looks something like this:

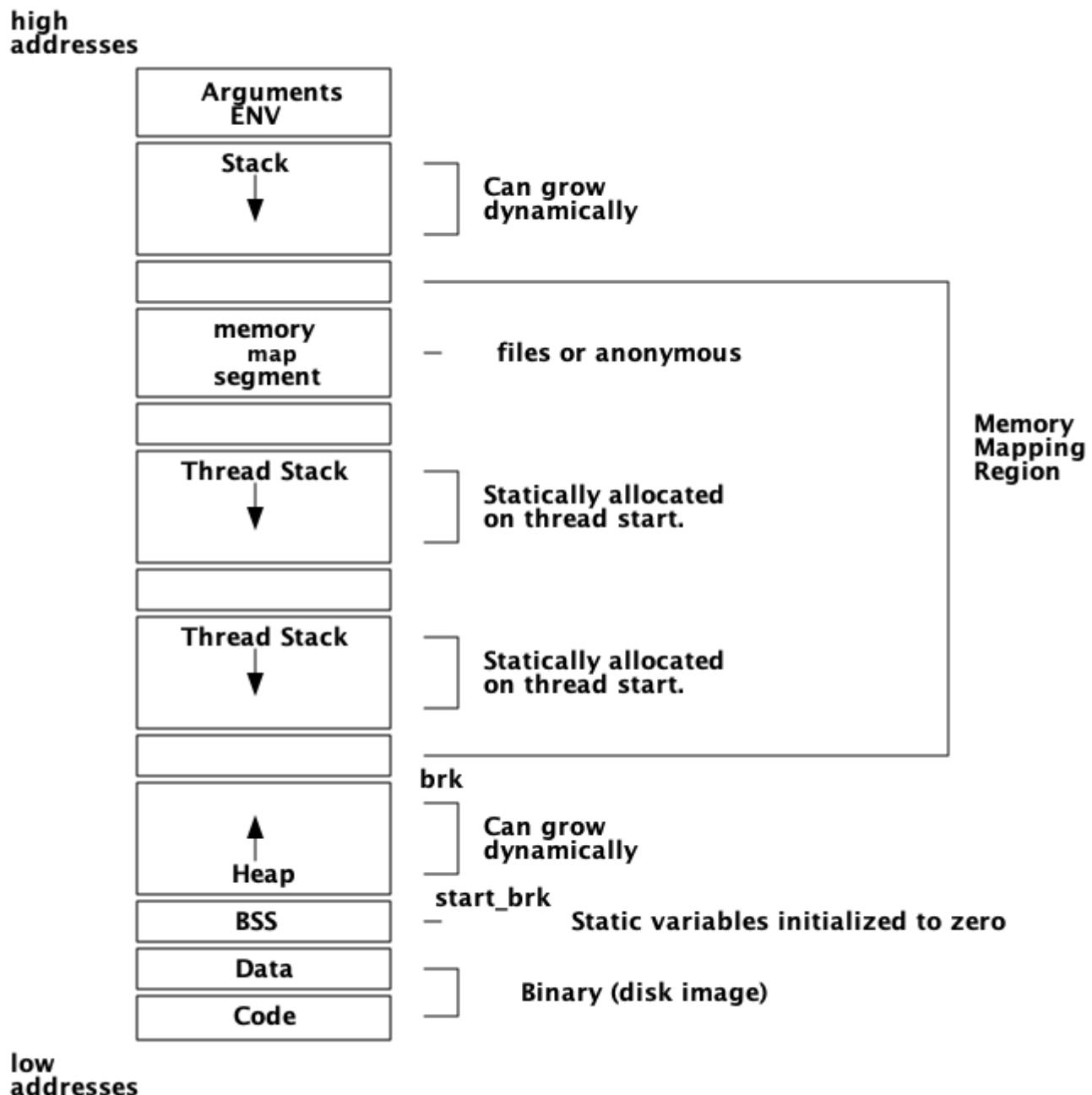


Figure 24. Program Memory Layout

Even though this picture might look daunting it is still a simplification. (For a full understanding of the memory subsystem read a book like "Understanding the Linux Kernel" or "Linux System Programming") What I want you to take away from this is that there are two types of dynamically allocatable memory: the heap and memory mapped segments. I will try to call this heap the *C-heap* from now on, to distinguish it from an Erlang process heap. I will call a memory mapped segment for just a *segment*, and any of the stacks in this picture for the *C-stack*.

The C-heap is allocated through malloc and a segment is allocated with mmap.

1. A note on pictures of memory

**Note** When drawing overview pictures of system memory and stacks we will follow the convention that memory addresses grows upward. That is low memory addresses on the bottom of the page and high memory addresses on the top of the page. (Stacks most often grow downward starting at high addresses, so that new elements are pushed at the lowest address.)

However when we draw a c-structure we will draw the fields from the top and down, even though the first field of the structure will be at the lowest address and the following fields at higher addresses. So pictures of structures have low address at the top of the page and high address at the bottom of the page.

This means that a picture of a c-structure and a picture of a memory area will have their address positions on the page mirrored. This becomes somewhat confusing when we try to picture structures and heaps in the same picture.

## 12.1. The memory subsystem

Now that we dive into the memory subsystem it will once again be apparent that ERTS is more like an operating system than just a programming language environment. Not only does ERTS provide a garbage collector for Erlang terms on the Erlang process level, but it also provides a plethora of low level memory allocators and memory allocation strategies.

For an overview of memory allocators see the erts\_alloc documentation at: [http://www.erlang.org/doc/man/erts\\_alloc.html](http://www.erlang.org/doc/man/erts_alloc.html)

All these allocators also comes with a number of parameters that can be used to tweak their behavior, and this is probably one of the most important areas from an operational point of view. This is where we can configure the system behavior to fit anything from a small embedded control system (like a Raspberry Pi) to an Internet scale 2TB database server.

There are currently eleven different allocators, six different allocation strategies, and more than 18 other different settings, some of which are taking arbitrary numerical values. This means that there basically is an infinite number of possible configurations. (OK, strictly speaking it is not infinite, since each number is bounded, but there are more configurations than you can shake a stick at.)

In order to be able to use these settings in any meaningful way we will have to understand how these allocators work and how each setting impacts the performance of the allocator.

The erts\_alloc manual goes as far as to give the following warning:



Only use these flags if you are absolutely sure what you are doing. Unsuitable settings may cause serious performance degradation and even a system crash at any time during operation.

— Ericsson AB, [http://www.erlang.org/doc/man/erts\\_alloc.html](http://www.erlang.org/doc/man/erts_alloc.html)

Making you absolutely sure that you know what you are doing, that is what this chapter is about.

Oh yes, we will also go into details of how the garbage collector works.

## 12.2. Different type of memory allocators

The Erlang run-time system is trying its best to handle memory in all situations and under all types of loads, but there are always corner cases. In this chapter we will look at the details of how memory is allocated and how the different allocators work. With this knowledge and some tools that we will look at later you should be able to detect and fix problems if your system ends up in one of these corner cases.

For a nice story about the troubles the system might get into and how to analyze and correct the behavior read Fred Hébert's essay "[Troubleshooting Down the Logplex Rabbit Hole](#)" [<https://blog.heroku.com/archives/2013/11/7/logplex-down-the-rabbit-hole>].

When we are talking about a memory allocator in this book we have a specific meaning in mind. Each memory allocator manage allocations and deallocations of memory of a certain type. Each allocator is intended for a specific type of data and is often specialized for one size of data.

Each memory allocator implements the allocator interface that can use different algorithms and settings for the actual memory allocation.

The goal with having different allocators is to reduce fragmentation, by grouping allocations of the same size, and to increase performance, by making frequent allocations cheap.

There are two special, fundamental or generic, memory allocator types *sys\_alloc* and *mseg\_alloc*, and nine specific allocators implemented through the *alloc\_util* framework.

In the following sections we will go though the different allocators, with a little detour into the general framework for allocators (*alloc\_util*).

Each allocator has several names used in the documentation and in the C code. See [List of memory allocators](#). for a short list of all allocators and their names. The C-name is used in the C-code to refer to the allocator. The Type-name is used in erl\_alloc.types to bind allocation types to an allocator. The Flag is the letter used for setting parameters of that allocator when starting Erlang.

*Table 1. List of memory allocators.*

Name	Description	C-name	Type-name	Flag
Basic allocator	malloc interface	sys_alloc	SYSTEM	Y
Memory segment allocator	mmap interface	mseg_alloc	-	M
Temporary allocator	Temporary allocations	temp_alloc	TEMPORARY	T
Heap allocator	Erlang heap data	eheap_alloc	EHEAP	H
Binary allocator	Binary data	binary_alloc	BINARY	B
ETS allocator	ETS data	ets_alloc	ETS	E
Driver allocator	Driver data	driver_alloc	DRIVER	R
Short lived allocator	Short lived memory	sl_alloc	SHORT_LIVED	S
Long lived allocator	Long lived memory	ll_alloc	LONG_LIVED	L
Fixed allocator	Fixed size data	fix_alloc	FIXED_SIZE	F
Standard allocator	For most other data	std_alloc	STANDARD	D

### 12.2.1. The basic allocator: sys\_alloc

The allocator *sys\_alloc* can not be disabled, and is basically a straight mapping to the underlying OS malloc implementation in libc.

If a specific allocator is disabled then *sys\_alloc* is used instead.

All specific allocators uses either *sys\_alloc* or *mseg\_alloc* to allocate memory from the operating system as needed.

When memory is allocated from the OS *sys\_alloc* can add (pad) a fixed number of kilobytes to the requested number. This can reduce the number of system calls by over allocating memory. The default padding is zero.

When memory is freed, *sys\_alloc* will keep some free memory allocated in the process. The size of this free memory is called the trim threshold, and the default is 128 kilobytes. This also reduces the number of system calls at the cost of a higher memory footprint. This means that if you are running the system with the default settings you can experience that the Beam process does not give memory back to the OS directly as memory is freed up.

Memory areas allocated by *sys\_alloc* are stored in the C-heap of the beam process which will grow as needed through system calls to brk.

## 12.2.2. The memory segment allocator: mseg\_alloc

If the underlying operating system supports mmap a specific memory allocator can use mseg\_alloc instead of sys\_alloc to allocate memory from the operating system.

Memory areas allocated through mseg\_alloc are called segments. When a segment is freed it is not immediately returned to the OS, instead it is kept in a segment cache.

When a new segment is allocated a cached segment is reused if possible, i.e. if it is the same size or larger than the requested size but not too large. The value of *absolute max cache bad fit* determines the number of kilobytes of extra size which is considered not too large. The default is 4096 kilobytes.

In order not to reuse a 4096 kilobyte segment for really small allocations there is also a *relative\_max\_cache\_bad\_fit* value which states that a cached segment may not be used if it is more than that many percent larger. The default value is 20 percent. That is a 12 KB segment may be used when asked for a 10 KB segment.

The number of entries in the cache defaults to 10 but can be set to any value from zero to thirty.

## 12.2.3. The memory allocator framework: alloc\_util

Building on top of the two generic allocators (sys\_alloc and mseg\_alloc) is a framework called *alloc\_util* which is used to implement specific memory allocators for different types of usage and data.

The framework is implemented in *erl\_alloc\_util.[ch]* and the different allocators used by ERTS are defined in erl\_alloc.types in the directory "erts/emulator/beam/".

In a SMP system there is usually one allocator of each type per scheduler thread.

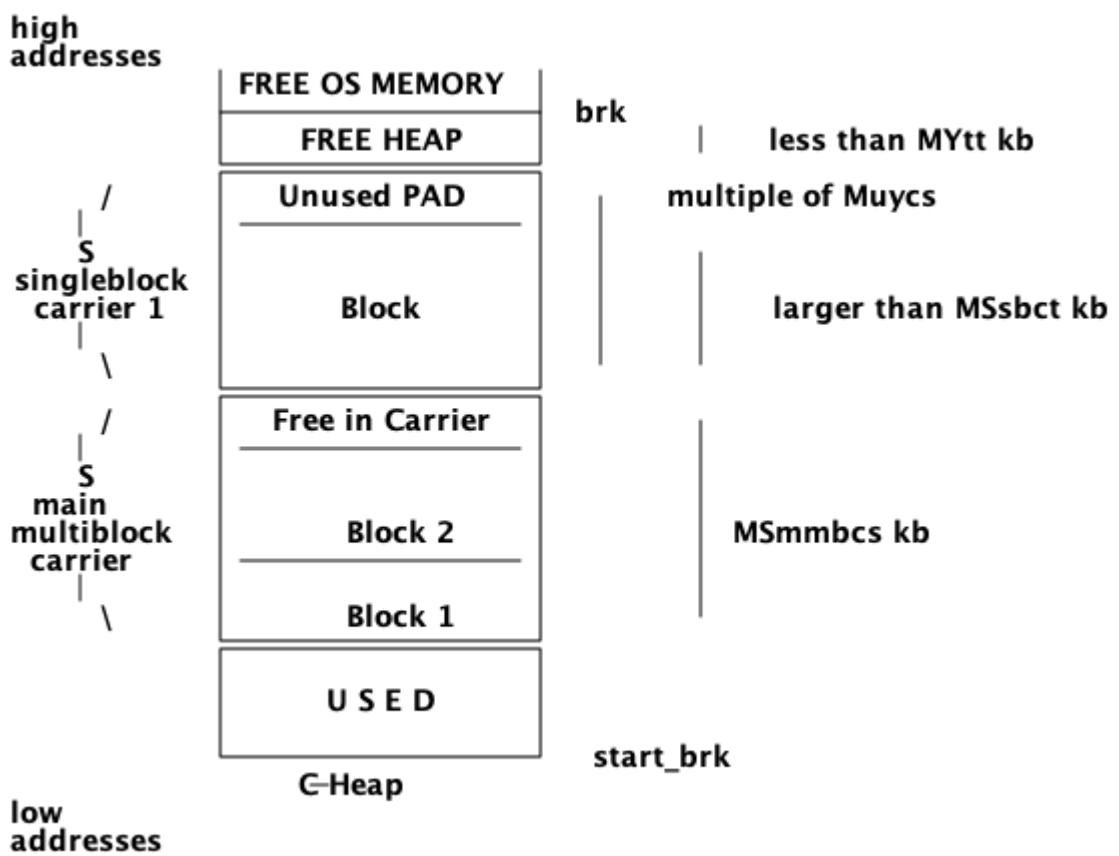
The smallest unit of memory that an allocator can work with is called a *block*. When you call an allocator to allocate a certain amount of memory what you get back is a block. It is also blocks that you give as an argument to the allocator when you want to deallocate memory.

The allocator does not allocate blocks from the operating system directly though. Instead the allocator allocates a *carrier* from the operating system, either through sys\_alloc or through mseg\_alloc, which in turn uses malloc or mmap. If sys\_alloc is used the carrier is placed on the C-heap and if mseg\_alloc is used the carrier is placed in a segment.

Small blocks are placed in a multiblock carrier. A multiblock carrier can as the name suggests contain many blocks. Larger blocks are placed in a singleblock carrier, which as the name implies contains one block.

What's considered a small and a large block is determined by the parameter *singleblock carrier threshold* (sbct), see the list of system flags below.

Most allocators also have one "main multiblock carrier" which is never deallocated.



## Memory allocation strategies

```
<!--
```

Are you intending for readers to take advantage of these allocation strategies in their code? If so, this section needs to be much more prominent, and not a subheading in a reference section. -  
bmacdonald

```
-->
```

To find a free block of memory in a multi block carrier an allocation strategy is used. Each type of allocator has a default allocation strategy, but you can also set the allocation strategy with the `as` flag.

The Erlang Run-Time System Application Reference Manual lists the following allocation strategies:

*Best fit:* Find the smallest block that satisfies the requested block size. (bf)

*Address order best fit:* Find the smallest block that satisfies the requested block size. If multiple blocks are found, choose the one with the lowest address. (aobf)

*Address order first fit:* Find the block with the lowest address that satisfies the requested block size. (aoff)

*Address order first fit carrier best fit* Find the carrier with the lowest address that can satisfy the requested block size, then find a block within that carrier using the "best fit" strategy. (aoffcbf)

*Address order first fit carrier address order best fit:* Find the carrier with the lowest address that can satisfy the requested block size, then find a block within that carrier using the "address order best fit" strategy. aoffcaobf (address order first fit carrier address order best fit)

*Good fit:* Try to find the best fit, but settle for the best fit found during a limited search. (gf)

*A fit:* Do not search for a fit, inspect only one free block to see if it satisfies the request. This strategy is only intended to be used for temporary allocations. (af)

— [erts\\_alloc](#) [[http://www.erlang.org/doc/man/erts\\_alloc.html](http://www.erlang.org/doc/man/erts_alloc.html)]

#### 12.2.4. The temporary allocator: temp\_alloc

The allocator *temp\_alloc*, is used for temporary allocations. That is very short lived allocations. Memory allocated by *temp\_alloc* may not be allocated over a Erlang process context switch.

You can use *temp\_alloc* as a small scratch or working area while doing some work within a function. Look at it as an extension of the C-stack and free it in the same way. That is, to be on the safe side, free memory allocated by *temp\_alloc* before returning from the function that did the allocation. There is a note in *erl\_alloc.types* saying that you should free a *temp\_alloc* block before the emulator starts executing Erlang code.

Note that no Erlang process running on the same scheduler as the allocator may start executing Erlang code before the block is freed. This means that you can not use a temporary allocation over a BIF or NIF trap (yield).

In a default R16 SMP system there is N+1 *temp\_alloc* allocators where N is the number of schedulers. The *temp\_alloc* uses the "A fit" (af) strategy. Since the allocation pattern of the *temp\_alloc* basically is that of a stack (mostly of size 0 or 1), this strategy works fine.

The temporary allocator is, in R16, used by the following types of data: TMP\_HEAP, MSG\_ROOTS, ROOTSET, LOADER\_TEMP, NC\_TMP, TMP, DCTRL\_BUF, TMP\_DIST\_BUF, ESTACK, DB\_TMP, DB\_MC\_STK, DB\_MS\_CMPL\_HEAP, LOGGER\_DSBUF, TMP\_DSBUF, DLL\_TMP\_BUF, TEMP\_TERM, SYS\_READ\_BUF, ENVIRONMENT, CON\_VPRINT\_BUF.

For an up to date list of allocation types allocated with each allocator, see *erl\_alloc.types* (e.g. grep TEMPORARY erts/emulator/beam/*erl\_alloc.types*).

I will not go through each of these different types, but in general as you can guess by their names, they are temporary buffers or work stacks.

#### 12.2.5. The heap allocator: eheap\_alloc

The heap allocator, is used for allocating memory blocks where tagged Erlang terms are stored, such as Erlang process heaps (all generations), heap fragments, and the beam\_registers.

This is probably the memory areas you are most interested in as an Erlang developer or when tuning an Erlang system. We will talk more about how these areas are managed in the upcoming sections on garbage collection and process memory. There we will also cover what a heap fragment is.

### 12.2.6. The binary allocator: `binary_alloc`

The binary allocator is used for, yes you guessed it, binaries. Binaries can be of quite varying sizes and have varying life spans. This allocator uses the *best fit* allocation strategy by default.

### 12.2.7. The ETS allocator: `ets_alloc`

The ETS allocator is used for most ETS related data, except for some short lived or temporary data used by ETS tables-

### 12.2.8. The driver allocator: `driver_alloc`

The driver allocator is used for ports, linked in drivers and NIFs.

### 12.2.9. The short lived allocator: `sl_alloc`

The short lived allocator is used for lists and buffers that are expected to be short lived. Short lived data can live longer than temporary data.

### 12.2.10. The long lived allocator: `ll_alloc`

The long lived allocator is used for long lived data, such as atoms, modules, funs and long lived tables

### 12.2.11. The fixed size allocator: `fix_alloc`

The fixed allocator is used for objects of a fixed size, such as PCBs, message refs and a few others. The fixed size allocator uses the *address order best fit* allocation strategy by default.

### 12.2.12. The standard allocator: `std_alloc`

The standard allocator is used by the other types of data. (active\_procs alloc\_info\_request arg\_reg bif\_timer\_ll bits\_buf bpd calls\_buf db\_heir\_data db\_heir\_data db\_named\_table\_entry dcache ddll\_handle ddll\_processes ddll\_processes dist\_entry dist\_tab driver\_lock ethread\_standard fd\_entry\_buf fun\_tab gc\_info\_request io\_queue line\_buf link\_lh module\_refs monitor\_lh monitor\_lh monitor\_sh nlink\_lh nlink\_lh nlink\_sh node\_entry node\_tab nodes\_monitor port\_data\_heap port\_lock port\_report\_exit port\_specific\_data proc\_dict process\_specific\_data ptimer\_ll re\_heap reg\_proc reg\_tab sched\_wall\_time\_request stack suspend\_monitor thr\_q\_element thr\_queue zlib )

## 12.3. TODO: system flags for memory

TODO

## 12.4. Process Memory

As we saw in [Processes](#) a process is really just a number of memory areas, in this chapter we will look a bit closer at how the stack, the heap and the mailbox are managed.

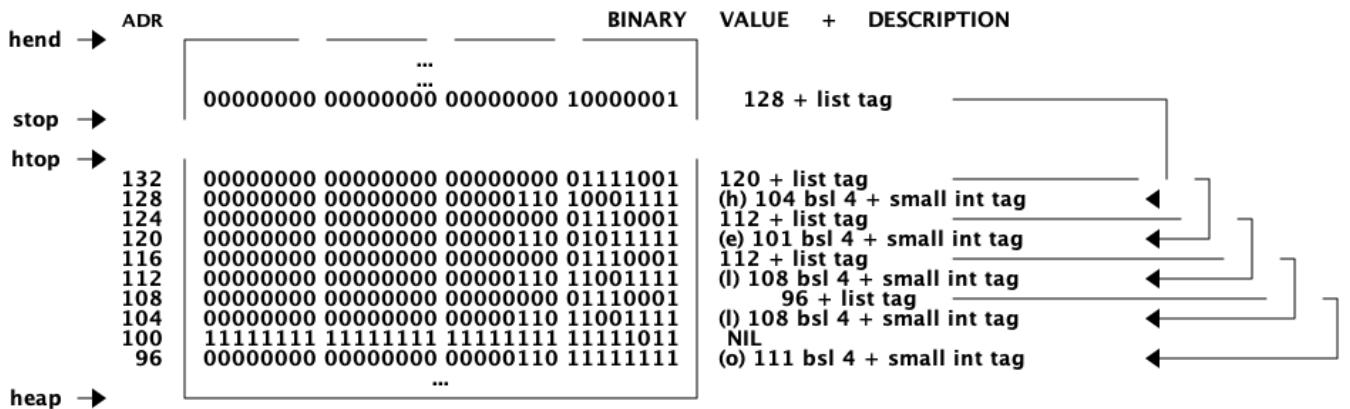
The default size of the stack and heap is 233 words. This default size can be changed globally when starting Erlang through the +h flag. You can also set the minimum heap size when starting a process with spawn\_opt by setting min\_heap\_size.

Erlang terms are tagged as we saw in [The Erlang Type System and Tags](#), and when they are stored on the heap they are either cons cells or boxed objects.

## 12.4.1. Term sharing

Objects on the heap are passed by references within the context of one process. If you call one function with a tuple as an argument, then only a tagged reference to that tuple is passed to the called function. When you build new terms you will also only use references to sub terms.

For example if you have the string "hello" (which is the same as this list of integers: [104,101,108,108,111]) you would get a stack layout similar to:



If you then create a tuple with two instances of the list, all that is repeated is the tagged pointer to the list: 000000000000000000000000000000001000001. The code

```
L = [104, 101, 108, 108, 111],  
T = {L, L}.
```

would result in a memory layout as seen below. That is, a boxed header saying that this is a tuple of size 2 and then two pointers to the same list.

ADR	VALUE	DESCRIPTION
144	000000000000000000000000000000001000001	128+CONS
140	000000000000000000000000000000001000001	128+CONS
136	000000000000000000000000000000001000000	2+ARITYVAL

This is nice, since it is cheap to do and uses very little space. But if you send the tuple to another process or do any other type of IO, or any operations which results in something called a *deep copy*, then the data structure is expanded. So if we send out tuple T to another process P2 (P2 ! T) then the heap of P2 will look like in:

```
..
```

You can quickly bring down your Erlang node by expanding a highly shared term, see [share.erl](#).

```

-module(share).

-export([share/2, size/0]).

share(0, Y) -> {Y,Y};
share(N, Y) -> [share(N-1, [N|Y]) || _ <- Y].


size() ->
    T = share:share(5,[a,b,c]),
    {{size, erts_debug:size(T)},
     {flat_size, erts_debug:flat_size(T)}}.

1> timer:tc(fun() -> share:share(10,[a,b,c]), ok end).
{1131,ok}

2> share:share(10,[a,b,c]), ok.
ok

3> byte_size(list_to_binary(test:share(10,[a,b,c]))), ok.
HUGE size (13695500364)
Abort trap: 6

```

You can calculate the memory size of a shared term and the size of the expanded size of the term with the functions `erts_debug:size/1` and `erts_debug:flat_size/1`.

```

> share:size().
{{size,19386},{flat_size,94110}}

```

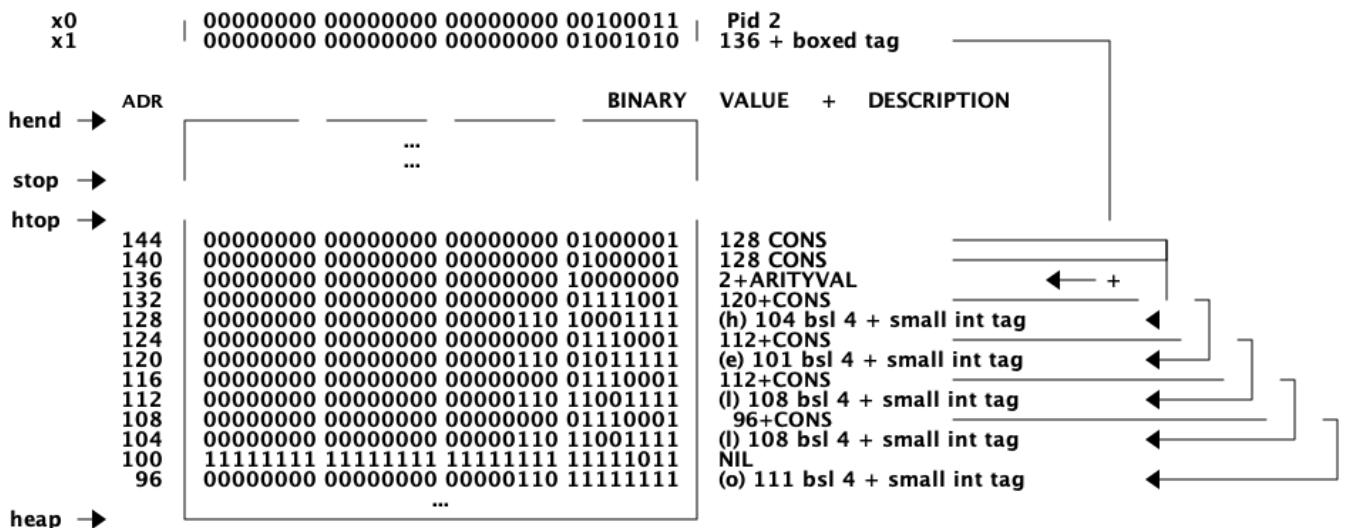
For most applications this is not a problem, but you should be aware of the problem, which can come up in many situations. A deep copy is used for IO, ETS tables, `binary_to_term`, and message passing.

Let us look in more detail how message passing works.

## 12.4.2. Message passing

When a process P1 sends a message M to another (local) process P2, the process P1 first calculates the flat size of M. Then it allocates a new message buffer of that size by doing a `heap_alloc` of a `heap_frag` in the local scheduler context.

Given the code in `send.erl` the state of the system could look like this just before the send in `p1/1`:



P2

Then P1 starts sending the message M to P2. It (through the code in `erl_message.c`) first calculates the flat size of M (which in our example is 23 words)<sup>[2]</sup>. Then (in a SMP system) if it can take a lock on P2 and there is enough room on the heap of P2 it will copy the message to the heap of P2.

If P2 is running (or exiting) or there isn't enough space on the heap, then a new heap fragment is allocated (of `sizeof(ErlHeapFragment) - sizeof(Eterm) + 23*sizeof(Eterm)`)<sup>[3]</sup> which after initialization will look like:

```
erl_heap_fragment:
ErlHeapFragment* next;      NULL
ErlOffHeap off_heap;
erl_off_heap_header* first; NULL
Uint64 overhead;           0
unsigned alloc_size;        23
unsigned used_size;         23
Eterm mem[1];              ?
... 22 free words
```

Then the message is copied into the heap fragment:

```

erl_heap_fragment:
ErlHeapFragment* next;      NULL
ErlOffHeap off_heap:
    erl_off_heap_header* first; Boxed tag+&mem+2*WS-
    Uint64 overhead;          0
    unsigned alloc_size;      23
    unsigned used_size;       23
    Eterm mem:
        2+ARITYVAL   <-----+
        &mem+3*WS+1  ---+
        &mem+13*WS+1 -----+
        (H*16)+15   <-+ |
        &mem+5*WS+1  --+ |
        (e*16)+15   <-+ |
        &mem+7*WS+1  ----| |
        (l*16)+15   <-+ |
        &mem+9*WS+1  ---+ |
        (l*16)+15   <-+ |
        &mem+11*WS+1 ---+ |
        (o*16)+15   <-+ |
        NIL
        (H*16)+15   <-----+
        &mem+15*WS+1 --+
        (e*16)+15   <-+
        &mem+17*WS+1 ----|
        (l*16)+15   <-+
        &mem+19*WS+1 ---+
        (l*16)+15   <-+
        &mem+21*WS+1 ---+
        (o*16)+15   <-+
        NIL</pre>

```

In either case a new mbox (ErlMessage) is allocated, a lock (ERTS\_PROC\_LOCK\_MSGQ) is taken on the receiver and the message on the heap or in the new heap fragment is linked into the mbox.

```

erl_msg {
    struct erl_msg* next = NULL;
    data: ErlHeapFragment *heap_frag = bp;
    Eterm m[0]           = message;
} ErlMessage;

```

Then the mbox is linked into the in message queue (msg\_inq) of the receiver, and the lock is released. Note that msg\_inq.last points to the next field of the last message in the queue. When a new mbox is linked in this next pointer is updated to point to the new mbox, and the last pointer is updated to point to the next field of the new mbox.

### 12.4.3. Binaries

As we saw in [The Erlang Type System and Tags](#) there are four types of binaries internally. Three of these types, *heap binaries*, *sub binaries* and *match contexts* are stored on the local heap and handled by the garbage collector and message passing as any other object, copied as needed.

## Reference Counting

The fourth type, large binaries or *refc binaries* on the other hand are partially stored outside of the process heap and they are reference counted.

The payload of a refc binary is stored in memory allocated by the binary allocator. There is also a small reference to the payload call a ProcBin which is stored on the process heap. This reference is copied by message passing and by the GC, but the payload is untouched. This makes it relatively cheap to send large binaries to other processes since the whole binary doesn't need to be copied.

All references through a ProcBin to a refc binary increases the reference count of the binary by one. All ProcBin objects on a process heap are linked together in a linked list. After a GC pass this linked list is traversed and the reference count of the binary is decreased with one for each ProcBin that has deceased. If the reference count of the refc binary reaches zero that binary is deallocated.

Having large binaries reference counted and not copied by send or garbage collection is a big win, but there is one problem with having a mixed environment of garbage collection and reference counting. In a pure reference counted implementation the reference count would be reduced as soon as a reference to the object dies, and when the reference count reaches zero the object is freed. In the ERTS mixed environment a reference to a reference counted object does not die until a garbage collection detects that the reference is dead.

This means that binaries, which has a tendency to be large or even huge, can hang around for a long time after all references to the binary are dead. Note that since binaries are allocated globally, all references from all processes need to be dead, that is all processes that has seen a binary need to do a GC.

Unfortunately it is not always easy, as a developer, to see which processes have seen a binary in the GC sense of the word seen. Imagine for example that you have a load balancer that receives work items and dispatches them to workers.

In [this code](#) there is an example of a loop which doesn't need to do GC. (See [listing 1b](#) for a full example.)

```
loop(Workers, N) ->
    receive
        WorkItem ->
            Worker = lists:nth(N+1, Workers),
            Worker ! WorkItem,
            loop(Workers, (N+1) rem length(Workers))
    end.
```

This server will just keep on grabbing references to binaries and never free them, eventually using up all system memory.

When one is aware of the problem it is easy to fix, one can either do a `garbage_collect` on each iteration of `loop` or one could do it every five seconds or so by adding an after clause to the receive. (*after 5000 → garbage\_collect()*, `loop(Workers, N)`).

## Sub Binaries and Matching

When you match out a part of a binary you get a sub binary. This sub binary will be a small structure just containing pointers into the real binary. This increases the reference count for the binary but uses very little extra space.

If a match would create a new copy of the matched part of the binary it would cost both space and time. So in most cases just doing a pattern match on a binary and getting a sub binary to work on is just what you want.

There are some degenerate cases, imagine for example that you load huge file like a book into memory and then you match out a small part like a chapter to work on. The problem is then that the whole of the rest of the book is still kept in memory until you are done with processing the chapter. If you do this for many books, perhaps you want to get the introduction of every book in your file system, then you will keep the whole of each book in memory and not just the introductory chapter. This might lead to huge memory usage.

The solution in this case, when you know you only want one small part of a large binary and you want to have the small part hanging around for some time, is to use `binary:copy/1`. This function is only used for its side effect, which is to actually copy the sub binary out of the real binary removing the reference to the larger binary and therefore hopefully letting it be garbage collected.

There is a pretty thorough explanation of how binary construction and matching is done in the Erlang documentation: [http://www.erlang.org/doc/efficiency\\_guide/binaryhandling.html](http://www.erlang.org/doc/efficiency_guide/binaryhandling.html).

#### 12.4.4. Garbage Collection

```
<!--  
This part of the content seems to be good, and probably worthy of being a top-level heading. It might  
be a bit long, though. - bmacdonald  
-->
```

When a process runs out of space on the stack and heap the process will try to reclaim space by doing a minor garbage collection. The code for this can be found in `erl_gc.c` [[https://github.com/erlang/otp/blob/maint/erts/emulator/beam/erl\\_gc.c](https://github.com/erlang/otp/blob/maint/erts/emulator/beam/erl_gc.c)].

ERTS uses a generational copying garbage collector. A copying collector means that during garbage collection all live young terms are copied from the old heap to a new heap. Then the old heap is discarded. A generational collector works on the principle that most terms die young, they are temporary terms created, used, and thrown away. Older terms are promoted to the old generation which is collected more seldom, with the rational that once a term has become old it will probably live for a long time.

Conceptually a garbage collection cycle works as follows:

- First you collect all roots (e.g. the stack).
- Then for each root, if the root points to a heap allocated object which doesn't have a forwarding pointer you copy the object to the new heap. For each copied object update the original with a forwarding pointer to the new copy.
- Now go through the new heap and do the same as for the roots.

We will go through an example to see how this is done in detail. We will go through a minor collection without an old generation, and we will only use the stack as the root set. In reality the process dictionary, trace data and probe data among other things are also included in the rootset.

Let us look at how the call to `garbage_collect` in the `gc_example` behaves. The code will generate a string which is shared by two elements of a cons and a tuple, the tuple will be eliminated resulting in garbage. After the GC there should only be one string on the heap. That is, first we generate the term `{"Hello", "Hello"}, "Hello"` (sharing the same string "Hello" in all instances. Then we just keep the term `["Hello", "Hello"]` when triggering a GC.



We will take the opportunity to go through how you, on a linux system, can use `gdb` to examine the behavior of ERTS. You can of course use the debugger of your choice. If you already know how to use `gdb` or if you have no interest in going into the debugger you can just ignore the meta text about how to inspect the system and just look at the diagrams and the explanations of how the GC works.

```

-module(gc_example).
-export([example/0]).

example() ->
    T = gen_data(),
    S = element(1, T),
    erlang:garbage_collect(),
    S.

gen_data() ->
    S = gen_string($H, $e, $l, $l, $o),
    T = gen_tuple([S,S],S),
    T.

gen_string(A,B,C,D,E) ->
    [A,B,C,D,E].

gen_tuple(A,B) ->
    {A,B}.

```

After compiling the example I start an erlang shell, test the call and prepare for a new call to the example (without hitting return):

```

1> gc_example:example().
["Hello","Hello"]
2> spawn(gc_example,example,[]).

```

Then I use gdb to attach to my erlang node (OS PID: 2955 in this case)

```
$ gdb /home/happi/otp/lib/erlang/erts-6.0/bin/beam.smp 2955
```



Depending on your settings for ptrace\_scope you might have to precede the gdb invocation with 'sudo'.

Then in gdb I set a breakpoint at the start of the main GC function and let the node continue:

```

(gdb) break garbage_collect_0
(gdb) cont
Continuing.

```

Now I hit enter in the Erlang shell and execution stops at the breakpoint:

```

Breakpoint 1, garbage_collect_0 (A__p=0x7f673d085f88, BIF__ARGS=0x7f673da90340) at beam/bif.c:3771
3771      FLAGS(BIF_P) |= F_NEED_FULLSWEEP;

```

Now we can inspect the PCB of the process:

```
(gdb) p *(Process *) A__p
$1 = {common = {id = 1408749273747, refc = {counter = 1}, tracer_proc = 18446744073709551611,
trace_flags = 0, u = {alive = {
    started_interval = 0, reg = 0x0, links = 0x0, monitors = 0x0, ptimer = 0x0}, release = {later
= 0, func = 0x0, data = 0x0,
    next = 0x0}}}, htop = 0x7f6737145950, stop = 0x7f6737146000, heap = 0x7f67371458c8, hend =
0x7f6737146010, heap_sz = 233,
min_heap_size = 233, min_vheap_size = 46422, fp_exception = 0, hipe = {nsp = 0x0, nstack = 0x0,
nstend = 0x0, ncallee = 0x7f673d080000,
closure = 0, nstgraylim = 0x0, nstblacklim = 0x0, ngra = 0x0, ncsp = 0x7f673d0863e8, arity = 0,
float_result = 0}, arity = 0,
arg_reg = 0x7f673d086080, max_arg_reg = 6, def_arg_reg = {393227, 457419, 18446744073709551611,
233, 46422, 2000}, cp = 0x7f673686ac40,
i = 0x7f673be17748, catches = 0, fcalls = 1994, rcount = 0, schedule_count = 0, reds = 0,
group_leader = 893353197987, flags = 0,
fvalue = 18446744073709551611, freason = 0, ftrace = 18446744073709551611, next = 0x7f673d084cc0,
nodes_monitors = 0x0,
suspend_monitors = 0x0, msg = {first = 0x0, last = 0x7f673d086120, save = 0x7f673d086120, len = 0,
mark = 0x0, saved_last = 0x7d0}, u = {
bif_timers = 0x0, terminate = 0x0}, dictionary = 0x0, seq_trace_clock = 0, seq_trace_lastcnt = 0,
seq_trace_token = 18446744073709551611, initial = {393227, 457419, 0}, current = 0x7f673be17730,
parent = 1133871366675,
approx_started = 1407857804, high_water = 0x7f67371458c8, old_hend = 0x0, old_htop = 0x0, old_heap =
0x0, gen_gcs = 0,
max_gen_gcs = 65535, off_heap = {first = 0x0, overhead = 0}, mbuf = 0x0, mbuf_sz = 0, psd = 0x0,
bin_vheap_sz = 46422,
bin_vheap_mature = 0, bin_old_vheap_sz = 46422, bin_old_vheap = 0, sys_task_qs = 0x0, state =
{counter = 41002}, msg_inq = {first = 0x0,
last = 0x7f673d086228, len = 0}, pending_exit = {reason = 0, bp = 0x0}, lock = {flags = {counter
= 1}, queue = {0x0, 0x0, 0x0, 0x0}},
refc = {counter = 1}}, scheduler_data = 0x7f673bd6c080, suspendee = 18446744073709551611,
pending_suspenders = 0x0, run_queue = {
counter = 140081362118912}, hipe_smp = {have_receive_locks = 0}}
```

Wow, that was a lot of information. The interesting part is about the stack and the heap:

```
hend = 0x7f6737146010,
stop = 0x7f6737146000,
htop = 0x7f6737145950,
heap = 0x7f67371458c8,
```

By using some helper scripts we can inspect the stack and the heap in a meaningful way. (see [Full Code Listings](#) for the definitions of the scripts in `gdb_script`.)

```
(gdb) source gdb_scripts
(gdb) print_p_stack A__p
0x00007f6737146008 [0x00007f6737145929] cons -> 0x00007f6737145928
(gdb) print_p_heap A__p
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f6737145919] cons -> 0x00007f6737145918
0x00007f6737145928 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145920 [0xfffffffffffffb] NIL
0x00007f6737145918 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145910 [0x00007f67371458f9] cons -> 0x00007f67371458f8
0x00007f6737145908 [0x00000000000048f] 72
0x00007f6737145900 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371458f8 [0x00000000000065f] 101
0x00007f67371458f0 [0x00007f67371458d9] cons -> 0x00007f67371458d8
0x00007f67371458e8 [0x0000000000006cf] 108
0x00007f67371458e0 [0x00007f67371458c9] cons -> 0x00007f67371458c8
0x00007f67371458d8 [0x0000000000006cf] 108
0x00007f67371458d0 [0xfffffffffffffb] NIL
0x00007f67371458c8 [0x0000000000006ff] 111
```

Here we can see the heap of the process after it has allocated the list "Hello" on the heap and the cons containing that list twice, and the tuple containing the cons and the list. The *root set*, in this case the stack, contains a pointer to the cons containing two copies of the list. The tuple is dead, that is, there are no references to it.

The garbage collection starts by calculating the root set and by allocating a new heap (*to space*). By stepping into the GC code in the debugger you can see how this is done. I will not go through the details here. After a number of steps the execution will reach the point where all terms in the root set are copied to the new heap. This starts around (depending on version) line 1272 with a while loop in `erl_gc.c`.

In our case the root is a cons pointing to address 0x00007f95666597f0 containing the letter (integer) 'H'. When a cons cell is moved from the current heap, called *from space*, to *to space* the value in the head (or car) is overwritten with a *moved cons* tag (the value 0).

After the first step where the root set is moved, the *from space* and the *to space* looks like this:

from space:

```
(gdb) print_p_heap p
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f67371445b1] cons -> 0x00007f67371445b0
0x00007f6737145928 [0x0000000000000000] Tuple size 0
0x00007f6737145920 [0xfffffffffffffb] NIL
0x00007f6737145918 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145910 [0x00007f67371458f9] cons -> 0x00007f67371458f8
0x00007f6737145908 [0x00000000000048f] 72
0x00007f6737145900 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371458f8 [0x00000000000065f] 101
0x00007f67371458f0 [0x00007f67371458d9] cons -> 0x00007f67371458d8
0x00007f67371458e8 [0x0000000000006cf] 108
0x00007f67371458e0 [0x00007f67371458c9] cons -> 0x00007f67371458c8
0x00007f67371458d8 [0x0000000000006cf] 108
0x00007f67371458d0 [0xfffffffffffffb] NIL
0x00007f67371458c8 [0x0000000000006ff] 111
```

to space:

```
(gdb) print_heap n_htop-1 n_htop-2
0x00007f67371445b8 [0x00007f6737145919] cons -> 0x00007f6737145918
0x00007f67371445b0 [0x00007f6737145909] cons -> 0x00007f6737145908
```

In *from space* the head of the first cons cell has been overwritten with 0 (looks like a tuple of size 0) and the tail has been overwritten with a forwarding pointer pointing to the new cons cell in the *to space*. In *to space* we now have the first cons cell with two backward pointers to the head and the tail of the cons in the *from space*.

When the collector is done with the root set the *to space* contains backward pointers to all still live terms. At this point the collector starts sweeping the *to space*. It uses two pointers *n\_hp* pointing to the bottom of the unseen heap and *n\_htop* pointing to the top of the heap.

```
n_htop:
0x00007f67371445b8 [0x00007f6737145919] cons -> 0x00007f6737145918
n_hp   0x00007f67371445b0 [0x00007f6737145909] cons -> 0x00007f6737145908
```

The GC will then look at the value pointed to by *n\_hp*, in this case a cons pointing back to the *from space*. So it moves that cons to the *to space*, incrementing *n\_htop* to make room for the new cons, and incrementing *n\_hp* to indicate that the first cons is seen.

from space:

```
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f67371445b1] cons -> 0x00007f67371445b0
0x00007f6737145928 [0x0000000000000000] Tuple size 0
0x00007f6737145920 [0xfffffffffffffb] NIL
0x00007f6737145918 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145910 [0x00007f67371445c1] cons -> 0x00007f67371445c0
0x00007f6737145908 [0x0000000000000000] Tuple size 0
0x00007f6737145900 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371458f8 [0x00000000000065f] 101
0x00007f67371458f0 [0x00007f67371458d9] cons -> 0x00007f67371458d8
0x00007f67371458e8 [0x0000000000006cf] 108
0x00007f67371458e0 [0x00007f67371458c9] cons -> 0x00007f67371458c8
0x00007f67371458d8 [0x0000000000006cf] 108
0x00007f67371458d0 [0xfffffffffffffb] NIL
0x00007f67371458c8 [0x0000000000006ff] 111
```

to space:

n\_htop:

```
0x00007f67371445c8 [0x00007f67371458f9] cons -> 0x00007f67371458f8
0x00007f67371445c0 [0x00000000000048f] 72
```

n_hp	0x00007f67371445b8 [0x00007f6737145919] cons -> 0x00007f6737145918
SEEN	0x00007f67371445b0 [0x00007f67371445c1] cons -> 0x00007f67371445c0

The same thing then happens with the second cons.

from space:

```
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f67371445b1] cons -> 0x00007f67371445b0
0x00007f6737145928 [0x0000000000000000] Tuple size 0
0x00007f6737145920 [0x00007f67371445d1] cons -> 0x00007f67371445d0
0x00007f6737145918 [0x0000000000000000] Tuple size 0
0x00007f6737145910 [0x00007f67371445c1] cons -> 0x00007f67371445c0
0x00007f6737145908 [0x0000000000000000] Tuple size 0
0x00007f6737145900 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371458f8 [0x000000000000065f] 101
0x00007f67371458f0 [0x00007f67371458d9] cons -> 0x00007f67371458d8
0x00007f67371458e8 [0x00000000000006cf] 108
0x00007f67371458e0 [0x00007f67371458c9] cons -> 0x00007f67371458c8
0x00007f67371458d8 [0x00000000000006cf] 108
0x00007f67371458d0 [0xfffffffffffffb] NIL
0x00007f67371458c8 [0x00000000000006ff] 111
```

to space:

n\_htop:

0x00007f67371445d8	[0xfffffffffffffb]	NIL
0x00007f67371445d0	[0x00007f6737145909]	cons -> 0x00007f6737145908
0x00007f67371445c8	[0x00007f67371458f9]	cons -> 0x00007f67371458f8
n_hp	0x00007f67371445c0	[0x00000000000048f] 72
SEEN	0x00007f67371445b8	[0x00007f6737145919] cons -> 0x00007f67371445d0
SEEN	0x00007f67371445b0	[0x00007f67371445c1] cons -> 0x00007f67371445c0

The next element in *to space* is the immediate 72, which is only stepped over (with `n_hp++`). Then there is another cons which is moved.

The same thing then happens with the second cons.

from space:

```
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f67371445b1] cons -> 0x00007f67371445b0
0x00007f6737145928 [0x0000000000000000] Tuple size 0
0x00007f6737145920 [0x00007f67371445d1] cons -> 0x00007f67371445d0
0x00007f6737145918 [0x0000000000000000] Tuple size 0
0x00007f6737145910 [0x00007f67371445c1] cons -> 0x00007f67371445c0
0x00007f6737145908 [0x0000000000000000] Tuple size 0
0x00007f6737145900 [0x00007f67371445e1] cons -> 0x00007f67371445e0
0x00007f67371458f8 [0x0000000000000000] Tuple size 0
0x00007f67371458f0 [0x00007f67371458d9] cons -> 0x00007f67371458d8
0x00007f67371458e8 [0x0000000000006cf] 108
0x00007f67371458e0 [0x00007f67371458c9] cons -> 0x00007f67371458c8
0x00007f67371458d8 [0x0000000000006cf] 108
0x00007f67371458d0 [0xfffffffffffffb] NIL
0x00007f67371458c8 [0x0000000000006ff] 111
```

to space:

n\_htop:

```
0x00007f67371445e8 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371445e0 [0x00000000000065f] 101
0x00007f67371445d8 [0xfffffffffffffb] NIL
n_hp 0x00007f67371445d0 [0x00007f6737145909] cons -> 0x00007f6737145908
SEEN 0x00007f67371445c8 [0x00007f67371458f9] cons -> 0x00007f67371445e0
SEEN 0x00007f67371445c0 [0x00000000000048f] 72
SEEN 0x00007f67371445b8 [0x00007f6737145919] cons -> 0x00007f67371445d0
SEEN 0x00007f67371445b0 [0x00007f67371445c1] cons -> 0x00007f67371445c0
```

Now we come to a cons that points to a cell that has already been moved. The GC sees the IS\_MOVED\_CONS tag at 0x00007f6737145908 and copies the destination of the moved cell from the tail (`*n_hp++ = ptr[1];`). This way sharing is preserved during GC. This step does not affect *from space*, but the backward pointer in *to space* is rewritten.

to space:

n\_htop:

```
0x00007f67371445e8 [0x00007f67371458e9] cons -> 0x00007f67371458e8
0x00007f67371445e0 [0x00000000000065f] 101
n_hp 0x00007f67371445d8 [0xfffffffffffffb] NIL
SEEN 0x00007f67371445d0 [0x00007f67371445c1] cons -> 0x00007f67371445c0
SEEN 0x00007f67371445c8 [0x00007f67371458f9] cons -> 0x00007f67371445e0
SEEN 0x00007f67371445c0 [0x00000000000048f] 72
SEEN 0x00007f67371445b8 [0x00007f6737145919] cons -> 0x00007f67371445d0
SEEN 0x00007f67371445b0 [0x00007f67371445c1] cons -> 0x00007f67371445c0
```

Then the rest of the list (the string) is moved.

from space:

```
0x00007f6737145948 [0x00007f6737145909] cons -> 0x00007f6737145908
0x00007f6737145940 [0x00007f6737145929] cons -> 0x00007f6737145928
0x00007f6737145938 [0x0000000000000080] Tuple size 2
0x00007f6737145930 [0x00007f67371445b1] cons -> 0x00007f67371445b0
0x00007f6737145928 [0x0000000000000000] Tuple size 0
0x00007f6737145920 [0x00007f67371445d1] cons -> 0x00007f67371445d0
0x00007f6737145918 [0x0000000000000000] Tuple size 0
0x00007f6737145910 [0x00007f67371445c1] cons -> 0x00007f67371445c0
0x00007f6737145908 [0x0000000000000000] Tuple size 0
0x00007f6737145900 [0x00007f67371445e1] cons -> 0x00007f67371445e0
0x00007f67371458f8 [0x0000000000000000] Tuple size 0
0x00007f67371458f0 [0x00007f67371445f1] cons -> 0x00007f67371445f0
0x00007f67371458e8 [0x0000000000000000] Tuple size 0
0x00007f67371458e0 [0x00007f6737144601] cons -> 0x00007f6737144600
0x00007f67371458d8 [0x0000000000000000] Tuple size 0
0x00007f67371458d0 [0x00007f6737144611] cons -> 0x00007f6737144610
0x00007f67371458c8 [0x0000000000000000] Tuple size 0
```

to space:

n\_htop:

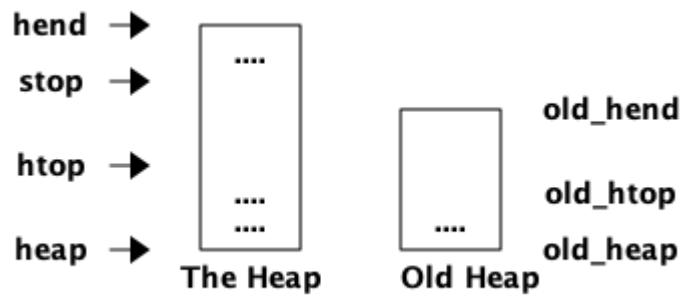
```
n_hp
SEEN 0x00007f6737144618 [0xfffffffffffffb] NIL
SEEN 0x00007f6737144610 [0x0000000000006ff] 111
SEEN 0x00007f6737144608 [0x00007f6737144611] cons -> 0x00007f6737144610
SEEN 0x00007f6737144600 [0x0000000000006cf] 108
SEEN 0x00007f67371445f8 [0x00007f6737144601] cons -> 0x00007f6737144600
SEEN 0x00007f67371445f0 [0x0000000000006cf] 108
SEEN 0x00007f67371445e8 [0x00007f67371445f1] cons -> 0x00007f67371445f0
SEEN 0x00007f67371445e0 [0x00000000000065f] 101
SEEN 0x00007f67371445d8 [0xfffffffffffffb] NIL
SEEN 0x00007f67371445d0 [0x00007f67371445c1] cons -> 0x00007f67371445c0
SEEN 0x00007f67371445c8 [0x00007f67371445e1] cons -> 0x00007f67371445e0
SEEN 0x00007f67371445c0 [0x00000000000048f] 72
SEEN 0x00007f67371445b8 [0x00007f67371445d1] cons -> 0x00007f67371445d0
SEEN 0x00007f67371445b0 [0x00007f67371445c1] cons -> 0x00007f67371445c0
```

There are some things to note from this example. When terms are created in Erlang they are created bottom up, starting with the elements. The garbage collector works top down, starting with the top level structure and then copying the elements. This means that the direction of the pointers change after the first GC. This has no real implications but it is good to know when looking at actual heaps. You can not assume that structures should be bottom up.

Also note that the GC does a breath first traversal. This means that locality for one term most often is worse after a GC. With the size of modern caches this should not be a problem. You could of course create a pathological example where it becomes a problem, but you can also create a pathological example where a depth first approach would cause problems.

The third thing to note is that sharing is preserved which is really important otherwise we might end up using more space after a GC than before.

Generations..



```
+high_water, old_hend, old_htop, old_heap,
gen_gcs, max_gen_gcs, off_heap, mbuf, mbuf_sz, psd, bin_vheap_sz,
bin_vheap_mature, bin_old_vheap_sz, bin_old_vheap+.
```

## 12.5. Other interesting memory areas

### 12.5.1. The atom table.

TODO ===== Code TODO ===== Constants TODO

[2] We ignore tracing here which will add a trace token to the size of the message, and always use a heap fragment.

[3] The -sizeof(Eterm) comes from mem in ErlHeapFragment already having the size of 1 Eterm



## Chapter 13

---

# Advanced data structures (ETS, DETS, Mnesia)



**Chapter 14****IO, Ports and Networking (10p)**

Within Erlang, all communication is done by asynchronous signaling. The communication between an Erlang node and the outside world is done through a *port*. A port is an interface between Erlang processes and an external resource. In early versions of Erlang a port behaved very much in the same way as a process and you communicated by sending and receiving signals. You can still communicate with ports this way but there are also a number of BIFs to communicate directly with a port.

In this chapter we will look at how ports are used as a common interface for all IO, how ports communicate with the outside world and how Erlang processes communicate with ports. But first we will look at how standard IO works on a higher level.

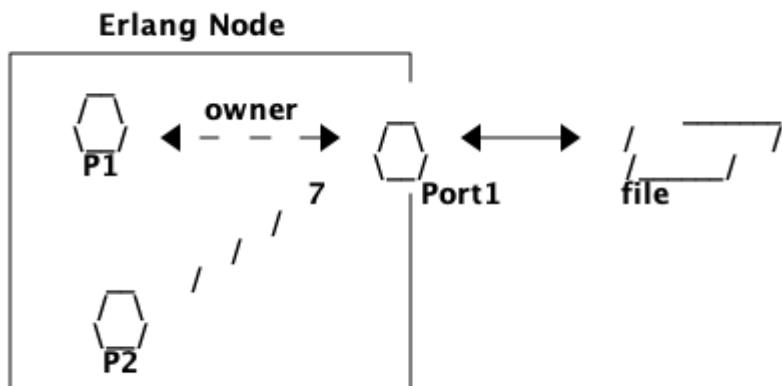
## 14.1. Standard IO

- IO protocol
- group leader
- erlang:display — a BIF that sends directly to node's std out
- io:format — sends through io protocol and the group leader
- Redirecting standard IO at startup (detached mode)
- Standard in and out

## 14.2. Ports

A port is the process like interface between Erlang processes and everything that is not Erlang processes. The programmer can to a large extent pretend that everything in the world behaves like an Erlang process and communicate through message passing.

Each port has an owner, more on this later, but all processes who know about the port can send messages to the port. In figure REF we see how a process can communicate with the port and how the port is communicating to the world outside the Erlang node.



*Figure 25. Port Communication*

Process P1 has opened a port (Port1) to a file, and is the owner of the port and can receive messages from the port. Process P2 also has a handle to the port and can send messages to the port. The processes and the port reside in an Erlang node. The file lives in the file and operating system on the outside of the Erlang node.

If the port owner dies or is terminated the port is also killed. When a port terminates all external resources should also be cleaned up. This is true for all ports that come with Erlang and if you implement your own port you should make sure it does this cleanup.

### 14.2.1. Different types of Ports

There are three different classes of ports: file descriptors, external programs and drivers. A file descriptor port makes it possible for a process to access an already opened file descriptor. A port to an external program invokes the external program as a separate OS process. A driver port requires a driver to be loaded in the Erlang node.

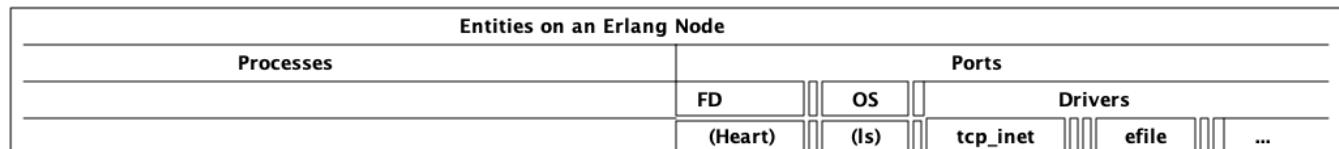
All ports are created by a call to `erlang:open_port(PortName, PortSettings)`.

A file descriptor port is opened with `{fd, In, Out}` as the PortName. This class of ports is used by some internal ERTS servers like the old shell. They are considered to not be very efficient and hence seldom used.

An external program port can be used to execute any program in the native OS of the Erlang node. To open an external program port you give either the argument `{spawn, Command}` or `{spawn_executable, FileName}` with the name of the external program. This is the easiest and one of the safest way to interact with code written in other programming languages. Since the external program is executed in its own OS process it will not bring down the Erlang node if it crashes. (It can of course use up all CPU or memory or do a number of other things to bring down the whole OS, but it is much safer than a linked in driver or a NIF).

A driver port requires that a driver program has been loaded with ERTS. Such a port is started with either `{spawn, Command}` or `{spawn_driver, Command}`. Writing your own linked in driver can be an efficient way to interface for example some C library code that you would like to use. Note that a linked in driver executes in the same OS process as the Erlang node and a crash in the driver will bring down the whole node. Details about how to write an Erlang driver in general can be found in [Interfacing C — BIFs NIFs and Linked in Drivers](#).

Erlang/OTP comes with a number port drivers implementing the predefined port types. There are the common drivers available on all platforms: `tcp_inet`, `udp_inet`, `sctp_inet`, `efile`, `zlib_drv`, `ram_file_drv`, `binary_filer`, `tty_sl`. These drivers are used to implement e.g. file handling and sockets in Erlang. On Windows there is also a driver to access the registry: `registry_drv`. And on most platforms there are example drivers to use when implementing your own driver like: `multi_drv` and `sig_drv`.



*Figure 26. Entities on an Erlang Node*

#### Ports to file descriptors

Creating

Commands

Examples

Implementation

#### Ports to Spawed OS Processes

Creating

Commands

Implementation

Windows

## Ports to Linked in Drivers

Creating

Commands

Implementation

See chapter [Interfacing C — BIFs NIFs and Linked in Drivers](#) for details of how to implement your own linked in driver.

## 14.3. Distributed Erlang

## 14.4. Sockets, UDP and TCP



## Chapter 15

---

# Distribution



## Chapter 16

---

# Interfacing C — BIFs NIFs and Linked in Drivers



## Chapter 17

# Native Code



# **Part II**

# **Running ERTS**



# Chapter 18

# Tracing



# Chapter 19

# Debugging



This chapter is still a stub and it's being heavily worked on. If planning a major addition to this chapter, please synchronize with the authors to avoid conflicts or duplicated efforts. You are still welcome to submit your feedback using a GitHub Issue. You can use the same mechanism to suggest sections that you believe should be included in this chapter, too.

## 19.1. Preliminary Outline

- Introduction
- debugger
- dbg
- redbug
- Crash dumps
- ...

## 19.2. Introduction

Debugging is the *art* of identifying and removing errors (i.e. *bugs*) from software. This section covers the most common Erlang debugging tools and techniques. Even if step-by-step debugging tools such as the [Debugger](http://erlang.org/doc/apps/debugger/debugger_chapter.html) [http://erlang.org/doc/apps/debugger/debugger\_chapter.html] exist in Erlang, the most effective debugging techniques in Erlang are the ones based on the so-called Erlang *tracing facilities*, which will be discussed in detail in chapter [Tracing](#). This chapter also covers the concept of *Crash Dump*, a readable text file generated by the Erlang Runtime System when an unrecoverable error is detected, for example when the system runs out of memory or when an emulator limit is reached. *Crash Dumps* can be extremely precious for post-mortem analysis of Erlang nodes and you will learn how to read and interpret them.

## 19.3. debugger

TODO

## 19.4. dbg

TODO

## 19.5. Redbug

*Redbug* is a debugging utility which allows you to easily interact with the Erlang *tracing facilities*. It is an external library and therefore it has to be installed separately. One of the best Redbug features is its ability to shut itself down in case of overload.

### 19.5.1. Installing Redbug

You can clone redbug via:

```
$ git clone https://github.com/massemanet/redbug
```

You can then compile it with:

```
$ cd redbug
$ make
```

Ensure `redbug` is included in your path when starting an Erlang shell and you are set to go. This can be done by explicitly adding the path to the `redbug beam` files when invoking `erl`:

```
$ erl -pa /path/to/redbug/ebin
```

Alternatively, the following line can be added to the `~/.erlang` file. This will ensure that the path to `redbug` gets included automatically at every startup:

```
code:add_patha("/path/to/redbug/ebin").
```

## 19.5.2. Using Redbug

`Redbug` is safe to be used in production, thanks to a self-protecting mechanism against overload, which kills the tool in case too many tracing messages are sent, preventing the Erlang node to become overloaded. Let's see it in action:

```
$ erl
Erlang/OTP 19 [erts-8.2] [...]

Eshell V8.2 (abort with ^G)
1> l(redbug). ①
{module,redbug}
2> redbug:start("lists:sort/1"). ②
{30,1}
3> lists:sort([3,2,1]).
[1,2,3]

% 15:20:20 <0.31.0>({erlang,apply,2}) ③
% lists:sort([3,2,1])
redbug done, timeout - 1 ④
```

① First, we ensure that the `redbug` module is available and loaded.

② We then start `redbug`. We are interested in the function named `sort` with arity `1`, exported by the module `lists`. Remember that, in Erlang lingo, the *arity* represents the number of input arguments that a given function takes.

③ Finally, we invoke the `lists:sort/1` function and we verify that a message is produced by `redbug`.

④ After the default timeout (15 seconds) is reached, `redbug` stops and displays the message "redbug done". `Redbug` is also kind enough to tell us the reason why it stopped (*timeout*) and the number of messages that collected until that point (*I*).

Let's now look at the actual message produced by `redbug`. By default messages are printed to the standard output, but it's also possible to dump them to file:

```
% 15:20:20 <0.31.0>({erlang,apply,2})
% lists:sort([3,2,1])
```

Depending on the version of redbug you are using, you may get a slightly different message. In this case, the message is split across two lines. The first line contains a **timestamp**, the **Process Identifier** (or *PID*) of the Erlang process which invoked the function and the **caller** function. The second line contains the function called, including the input arguments. Both lines are prepended with a %, which reminds us of the syntax for Erlang comments.

We can also ask Redbug to produce an extra message for the return value. This is achieved using the following syntax:

```
4> redbug:start("lists:sort/1->return").
{30,1}
```

Let's invoke the `lists:sort/1` function again. This time the output from redbug is slightly different.

```
5> lists:sort([3,2,1]).
[1,2,3]

% 15:35:52 <0.31.0>({erlang,apply,2})
% lists:sort([3,2,1])

% 15:35:52 <0.31.0>({erlang,apply,2})
% lists:sort/1 -> [1,2,3]
redbug done, timeout - 1
```

In this case two messages are produced, one when entering the function and one when leaving the same function.

When dealing with real code, trace messages can be complex and therefore hardly readable. Let's see what happens if we try to trace the sorting of a list containing 10.000 elements.

```

6> lists:sort(lists:seq(10000, 1, -1)).
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
23,24,25,26,27,28,29|...]

% 15:48:42.208 <0.77.0>({erlang,apply,2})
% lists:sort([10000,9999,9998,9997,9996,9995,9994,9993,9992,9991,9990,9989,9988,9987,9986,
% 9985,9984,9983,9982,9981,9980,9979,9978,9977,9976,9975,9974,9973,9972,9971,
% 9970,9969,9968,9967,9966,9965,9964,9963,9962,9961,9960,9959,9958,9957,9956,
% 9955,9954,9953,9952,9951,9950,9949,9948,9947,9946,9945,9944,9943,9942,9941,
% 9940,9939,9938,9937,9936,9935,9934,9933,9932,9931,9930,9929,9928,9927,9926,
% 9925,9924,9923,9922,9921,9920,9919,9918,9917,9916,9915,9914,9913,9912,9911,
% [...]
% 84,83,82,81,80,79,78,77,76,75,74,73,72,71,70,69,68,67,66,65,64,63,62,61,60,
% 59,58,57,56,55,54,53,52,51,50,49,48,47,46,45,44,43,42,41,40,39,38,37,36,35,
% 34,33,32,31,30,29,28,27,26,25,24,23,22,21,20,19,18,17,16,15,14,13,12,11,10,9,
% 8,7,6,5,4,3,2,1])

% 15:48:42.210 <0.77.0>({erlang,apply,2}) lists:sort/1 ->
% [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
% 23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,
% 42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,
% 61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,
% 80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,
% 99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,
% [...]
% 9951,9952,9953,9954,9955,9956,9957,9958,9959,9960,9961,
% 9962,9963,9964,9965,9966,9967,9968,9969,9970,9971,9972,
% 9973,9974,9975,9976,9977,9978,9979,9980,9981,9982,9983,
% 9984,9985,9986,9987,9988,9989,9990,9991,9992,9993,9994,
% 9995,9996,9997,9998,9999,10000]
redbug done, timeout - 1

```

Most of the output has been truncated here, but you should get the idea. To improve things, we can use a couple of redbug options. The option `{arity, true}` instructs redbug to only display the number of input arguments for the given function, instead of their actual value. The `{print_return, false}` option tells Redbug not to display the return value of the function call, and to display a `...` symbol, instead. Let's see these options in action.

```

7> redbug:start("lists:sort/1->return", [{arity, true}, {print_return, false}]).
{30,1}

8> lists:sort(lists:seq(10000, 1, -1)).
[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,
23,24,25,26,27,28,29|...]

% 15:55:32 <0.77.0>({erlang,apply,2})
% lists:sort/1

% 15:55:32 <0.77.0>({erlang,apply,2})
% lists:sort/1 -> '...'
redbug done, timeout - 1

```

By default, redbug stops after 15 seconds or after 10 messages are received. Those values are a safe default, but they are rarely

enough. You can bump those limits by using the `time` and `msgs` options. `time` is expressed in milliseconds.

```
9> redbug:start("lists:sort/1->return", [{arity, true}, {print_return, false}, {time, 60 * 1000}, {msgs, 100}]).  
{30,1}
```

We can also activate redbug for several function calls simultaneously. Let's enable tracing for both functions `lists:sort/1` and `lists:sort_1/3` (an internal function used by the former):

```
10> redbug:start(["lists:sort/1->return", "lists:sort_1/3->return"]).  
{30,2}  
  
11> lists:sort([4,4,2,1]).  
[1,2,4,4]  
  
% 18:39:26 <0.32.0>({erlang,apply,2})  
% lists:sort([4,4,2,1])  
  
% 18:39:26 <0.32.0>({erlang,apply,2})  
% lists:sort_1(4, [2,1], [4])  
  
% 18:39:26 <0.32.0>({erlang,apply,2})  
% lists:sort_1/3 -> [1,2,4,4]  
  
% 18:39:26 <0.32.0>({erlang,apply,2})  
% lists:sort/1 -> [1,2,4,4]  
redbug done, timeout - 2
```

Last but not least, redbug offers the ability to only display results for matching input arguments. This is when the syntax looks a bit like magic.

```
12> redbug:start(["lists:sort([1,2,5])->return"]).  
{30,1}  
  
13> lists:sort([4,4,2,1]).  
[1,2,4,4]  
  
14> lists:sort([1,2,5]).  
[1,2,5]  
  
% 18:45:27 <0.32.0>({erlang,apply,2})  
% lists:sort([1,2,5])  
  
% 18:45:27 <0.32.0>({erlang,apply,2})  
% lists:sort/1 -> [1,2,5]  
redbug done, timeout - 1
```

In the above example, we are telling redbug that we are only interested in function calls to the `lists:sort/1` function when the input arguments is the list `[1,2,5]`. This allows us to remove a huge amount of noise in the case our target function is used by many actors at the same time and we are only interested in a specific use case. Oh, and don't forget that you can use the underscore as a wildcard:

```
15> redbug:start(["lists:sort([1,_,5])->return"]). {30,1}

16> lists:sort([1,2,5]). [1,2,5]

% 18:49:07 <0.32.0>({erlang,apply,2}) lists:sort([1,2,5])

% 18:49:07 <0.32.0>({erlang,apply,2}) lists:sort/1 -> [1,2,5]

17> lists:sort([1,4,5]). [1,4,5]

% 18:49:09 <0.32.0>({erlang,apply,2}) lists:sort([1,4,5])

% 18:49:09 <0.32.0>({erlang,apply,2}) lists:sort/1 -> [1,4,5] rebug
% done, timeout - 2
```

This section does not pretend to be a comprehensive guide to rebug, but it should be enough to get you going. To get a full list of the available options for rebug, you can ask the tool itself:

```
18> redbug:help().
```

## 19.6. Crash Dumps

TODO

## Chapter 20

# Operation and Maintenance

One guiding principle behind the design of the runtime system is that bugs are more or less inevitable. Even if through an enormous effort you manage to build a bug free application you will soon learn that the world or your user changes and your application will need to be "fixed."

The Erlang runtime system is designed to facilitate change and to minimize the impact of bugs.

The impact of bugs is minimized by compartmentalization. This is done from the lowest level where each data structure is separate and immutable to the highest level where running systems are divided into separate nodes. Change is facilitated by making it easy to upgrade code and interacting and examining a running system.

## 20.1. Connecting to the System

We will look at many different ways to monitor and maintain a running system. There are many tools and techniques available but we must not forget the most basic tool, the shell and the ability to connect a shell to node.

In order to connect two nodes they need to share or know a secret pass phrase, called a cookie. As long as you are running both nodes on the same machine and the same user starts them they will automatically share the cookie (in the file `$HOME/.erlang.cookie`).

We can see this in action by starting two nodes, one Erlang node and one Elixir node. First we start an Erlang node called `node1`.

```
$ erl -sname node1
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
  [async-threads:10] [hipe] [kernel-poll:false]

Eshell V8.1 (abort with AB)
(node1@GDC08)1> nodes().
[]
(node1@GDC08)2>
```

Then we start an Elixir node called `node2`:

```
$ iex --sname node2
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
  [async-threads:10] [hipe] [kernel-poll:false]

Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iex(node2@GDC08)1>
```

In Elixir we can connect the nodes by running the command `Node.connect` name. In Erlang you do this with `net_kernel:connect(Name)`. The node connection is bidirectional so you only need to run the command on one of the nodes.

```
iex(node2@GDC08)1> Node.connect :node1@GDC08
true
iex(node2@GDC08)2>
```

In the distributed case this is somewhat more complicated since we need to make sure that all nodes know or share the cookie. This can be done in three ways. You can set the cookie used when talking to a specific node, you can set the same cookie for all systems at start up with the `-set_cookie` parameter, or you can copy the file `.erlang.cookie` to the home directory of the user running the system on each machine.

The last alternative, to have the same cookie in the cookie file of each machine in the system is usually the best option since it makes it easy to connect to the nodes from a local OS shell. Just set up some secure way of logging in to the machine either through VPN or ssh. In the next section we will see how to then connect a shell to a running node.

Using the second option it might look like this:

```
happi@GDC08:~$ cat ~/.erlang.cookie
pepparkaka
happi@GDC08:~$ ssh gds01

happi@gds01:~$ erl -sname node3 -setcookie pepparkaka
Erlang/OTP 18 [erts-7.3] [source-d2a6d81] [64-bit] [smp:8:8]
[async-threads:10] [hipe] [kernel-poll:false]

Eshell V7.3 (abort with ^G)
(node3@gds01)1> net_kernel:connect('node1@GDC08').
true
(node3@gds01)2> nodes().
[node1@GDC08, node2@GDC08]
(node3@gds01)3>
```

 A Potential Problem with Different Cookies Note that the default for the Erlang distribution is to create a fully connected network. That is, all nodes are connected to all other nodes in the network. In the example, once node3 connects to node1 it also is connected to node2. If each node has its own cookie you will have to tell each node the cookies of each other node before you try to connect them. You can start up a node with the flag `-connect_all false` in order to prevent the system from trying to make a fully connected network. Alternatively, you can start a node as hidden with the flag `-hidden`, which makes node connections to that node non transitive.

Now that we know how to connect nodes, even on different machines, to each other, we can look at how to connect a shell to a node.

## 20.2. The Shell

The Elixir and the Erlang shells works much the same way as a shell or a terminal window on your computer, except that they give you a terminal window directly into your runtime system. This gives you an extremely powerful tool, a basically CLI with full access to the runtime. This is fantastic for operation and maintenance.

In this section we will look at different ways of connecting to a node through the shell and some of the shell's perhaps less known but more powerful features.

## 20.2.1. Configuring Your Shell

Both the Elixir shell and the Erlang shell can be configured to provide you with shortcuts for functions that you often use.

The Elixir shell will look for the file `.iex.exs` first in the local directory and then in the users home directory. The code in this file is executed in the shell process and all variable bindings will be available in the shell.

In this file you can configure aspects such as the syntax coloring and the size of the history. [See hexdocs for a full documentation.](<https://hexdocs.pm/iex/IEx.html#module-the-iex-exs-file>)

You can also execute arbitrary code in the shell context.

When the Erlang runtime system starts, it first interprets the code in the Erlang configuration file. The default location of this file is in the users home directory `~/.erlang`.

This file is usually used to load the user default settings for the shell by adding the line

```
code:load_abs("/home/happi/.config/erlang/user_default").
```

Replace `/home/happi/.config/erlang/` with the absolute path you want to use.

If you call a local function from the shell it will try to call this function first in the module `user_default` and then in the module `shell_default` (located in `stdlib`). This is how command such as `ls()` and `help()` are implemented.

## 20.2.2. Connecting a Shell to a Node

When running a production system you will want to start the nodes in daemon mode through `run_erl`. We will go through how to start a node and some of the best practices for deployment and running in production in [xxx](#ch.live). Fortunately, even when you have started a system in daemon mode, without a shell, you can connect a shell to the system. There are actually several ways to do that. Most of these methods rely on the normal distribution mechanisms and hence require that you have the same Erlang cookie on both machines as described in the previous section.

### Remote shell (Remsh)

The easiest and probably the most common way to connect to an Erlang node is by starting a named node that connects to the system node through a remote shell. This is done with the `erl` command line flag `-remsh Name`. Note that you need to start a named node in order to be able to connect to another node, so you also need the `-name` or `-sname` flag. Also, note that these are arguments to the Erlang runtime so if you are starting an Elixir shell you need to add an extra `-` to the flags, like this:

```
$ iex --sname node4 --remsh node2@GDC08
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
[async-threads:10] [hipe] [kernel-poll:false]

Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iex(node2@GDC08)1>
```

Another thing to note here is that in order to start a remote Elixir shell you need to have IEx running on that node. There is no problem to connect Elixir and Erlang nodes to each other as we saw in the previous section, but you need to have the code of the shell you want to run loaded on the node you connect to.

It is also worth noting that there is no security built into either the normal Erlang distribution or to the remote shell implementation. You do not want to have your system node exposed to the internet and you do not want to connect from your local machine to a node. The safest way is probably to have a VPN tunnel to your live environment and use ssh to connect a machine running one of your live nodes. Then you can connect to one of the nodes using `remsh`.

It is important to understand that there are actually two nodes involved when you start a remote shell. The local node, named `node4` in the previous example and the remote node `node2`. These nodes can be on the same machine or on different machines. The local node is always running on the machine on which you gave the `iex` or `erl` command. On the local node there is a process running the `tty` program which interacts with the terminal window. The actual shell process runs on the remote node. This means, first of all, that the code for the shell you want to run (i.e. `iex` or the Erlang shell) has to exist at the remote node. It also means that code is executed on the remote node. And it also means that any shell default settings are taken from the settings of the remote machine.

Imagine that we have the following `.erlang` file in our home directory on the machine GDC08.

```
code:load_abs("/home/happi/.config/erlang/user_default").
```

```
io:format("ERTS is starting in ~n",[os:cmd("pwd")]).
```

And the `<filename>user_default.erl</filename>` file looks like this:

```
-module(user_default).
```

```
-export([tt/0]).
```

```
tt() -> test.
```

Then we create two directories `~/example/dir1` and `~/example/dir2` and we put two different `.iex.exs` files in those directories.

```
IO.puts "iEx starting in " pwd() IO.puts "iEx starting on " IO.puts Node.self
```

```
IEx.configure( colors: [enabled: true], alive_prompt: [ "\e[G", "(%node)", "%prefix", "<d1>", ] ) >> IO.ANSI.format >>  
IO.chardata_to_string )
```

```
IO.puts "iEx starting in " pwd() IO.puts "iEx starting on " IO.puts Node.self
```

```
IEx.configure( colors: [enabled: true], alive_prompt: [ "\e[G", "(%node)", "%prefix", "<d2>", ] ) >> IO.ANSI.format >>  
IO.chardata_to_string )
```

Now if we start four different nodes from these directories we will see how the shell configurations are loaded.

```
GDC08:~/example/dir1$ iex --sname node1  
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit]  
[smp:4:4] [async-threads:10] [hipe] [kernel-poll:false]  
  
ERTS is starting in /home/happi/example/dir1  
on [node1@GDC08]  
Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)  
iEx starting in  
/home/happi/example/dir1  
iEx starting on  
node1@GDC08  
(node1@GDC08)iex<d1>
```

```
GDC08:~/example/dir2$ iex --sname node2
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit]
[smp:4:4] [async-threads:10] [hipe] [kernel-poll:false]
```

```
ERTS is starting in /home/happi/example/dir2
on [node2@GDC08]
Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iEx starting in
/home/happi/example/dir2
iEx starting on
node2@GDC08
(node2@GDC08)iex<d2>
```

```
GDC08:~/example/dir1$ iex --sname node3 --remsh node2@GDC08
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
[async-threads:10] [hipe] [kernel-poll:false]
```

```
ERTS is starting in /home/happi/example/dir1
on [node3@GDC08]
Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iEx starting in
/home/happi/example/dir2
iEx starting on
node2@GDC08
(node2@GDC08)iex<d2>
```

```
GDC08:~/example/dir2$ erl -sname node4
Erlang/OTP 19 [erts-8.1] [source-0567896] [64-bit] [smp:4:4]
[async-threads:10] [hipe] [kernel-poll:false]
```

```
ERTS is starting in /home/happi/example/dir2
on [node4@GDC08]
Eshell V8.1 (abort with ^G)
(node4@GDC08)1> tt().
test
(node4@GDC08)2>
```

The shell configuration is loaded from the node running the shell, as you can see from the previous examples. If we were to connect to a node on a different machine, these configurations would not be present.

You can actually change which node and shell you are connected to by going into job control mode.

## Job Control Mode

By pressing control+G (ctrl-G) you enter the job control mode (JCL). You are then greeted by another prompt:

```
User switch command
-->
```

By typing **h** (followed by enter) you get a help text with the available commands in JCL:

```
c [nn]          - connect to job
i [nn]          - interrupt job
k [nn]          - kill job
j               - list all jobs
s [shell]       - start local shell
r [node [shell]] - start remote shell
q               - quit erlang
? | h          - this message
```

The interesting command here is the **r** command which starts a remote shell. You can give it the name of the shell you want to run, which is needed if you want to start an Elixir shell, since the default is the standard Erlang shell. Once you have started a new job (i.e. a new shell) you need to connect to that job with the **c** command. You can also list all jobs with **j**.

```
(node2@GDC08)iex<d2>
User switch command
--> r node1@GDC08 'Elixir.IEx'
--> c
Interactive Elixir (1.4.0) - press Ctrl+C to exit (type h() ENTER for help)
iEx starting in
/home/happi/example/dir1
iEx starting on
node1@GDC08
```

See the [Erlang Shell manual](<http://erlang.org/doc/man/shell.html>) for a full description of JCL mode.

You can quit your session by typing **ctrl+G q [enter]**. This shuts down the local node. You do **not** want to quit with any of **q()**, **halt()**, **init:stop()**, or **System.halt**. All of these will bring down the remote node which seldom is what you want when you have connected to a live server. Instead use **ctrl+\**, **ctrl+c ctrl+c, ctrl+g q [enter]** or **ctrl+c a [enter]**.

If you do not want to use a remote shell, which requires you to have two instances of the Erlang runtime system running, there are actually two other ways to connect to a node. You can also connect either through a Unix pipe or directly through ssh, but both of these methods require that you have prepared the node you want to connect to by starting it in a special way or by starting an ssh server.

### Connecting through a Pipe

By starting the node through the command **run\_erl** you will get a named pipe for IO and you can attach a shell to that pipe without the need to start a whole new node. As we shall see in the next chapter there are some advantages to using **run\_erl** instead of just starting Erlang in daemon mode, such as not losing standard IO and standard error output.

The **run\_erl** command is only available on Unix-like operating systems that implement pipes. If you start your system with **run\_erl**, something like:

```
> run_erl -daemon log/erl_pipe log "erl -sname node1"
```

or

```
> run_erl -daemon log/iex_pipe log "iex --sname node2"
```

You can then attach to the system through the named pipe (the first argument to `run_erl`).

```
> to_erl dir1/iex_pipe
iex(node2@GDC08)1>
```

You can exit the shell by sending EOF (`ctrl+d`) and leave the system running in the background. Note that with `to_erl` the terminal is connected directly to the live node so if you exit with `ctrl-G q [enter]` you will bring down that node, probably not what you want.

The last method for connecting to the node is through ssh.

### Connecting through SSH

Erlang comes with a built in ssh server which you can start on your node and then connect to directly. The [documentation for the ssh module](<http://erlang.org/doc/man/ssh.html>) explains all the details. For a quick test all you need is a server key which you can generate with `ssh-keygen`:

```
> mkdir ~ssh-test/
> ssh-keygen -t rsa -f ~ssh-test/ssh_host_rsa_key
```

Then you start the ssh daemon on the Erlang node:

```
gds01> erl
Erlang/OTP 18 [erts-7.3] [source-d2a6d81] [64-bit] [smp:8:8]
[async-threads:10] [hipe] [kernel-poll:false]

Eshell V7.3 (abort with ^G)
1> ssh:start().
{ok,<0.47.0>}
2> ssh:daemon(8021, [{system_dir, "/home/happi/.ssh/ehost/"},
                      {auth_methods, "password"}, {password, "pwd"}]).
```

You can now connect from another machine:

```
happi@GDC08:~> ssh -p 8021 happi@gds01
happi@gds01's password: [pwd]
Eshell V7.3 (abort with ^G)
1>
```

In a real world setting you would want to set up your server and user ssh keys as described in the documentation. At least you would want to have a better password.

To disconnect from the shell you need to shut down your terminal window. Using `q()` or `init:stop()` would bring down the node. In this shell you do not have access to neither JCL mode (`ctrl+g`) nor the BREAK mode (`ctrl+c`).

The break mode is really powerful when developing, profiling and debugging. We will take a look at it next.

### 20.2.3. Breaking (out or in).

When you press **ctrl+c** you enter BREAK mode. This is most often used just to break out of the shell by either typing **a [enter]** for abort or by hitting **ctrl+c** once more. But you can actually use this mode to break in to the internals of the Erlang runtime system.

When you enter BREAK mode you get a short menu:

```
BREAK: (a)abort (c)ontinue (p)roc info (i)nfo (l)oaded  
      (v)ersion (k)ill (D)b-tables (d)istribution
```

Abort exits the node and continue takes you back in to the shell. Hitting **p [enter]** will give you internal information about all processes in the system. We will look closer at what this information means in the next chapter (See [xxx](#ch.processes)).

You can also get information about the memory and the memory allocators in the node through the info choice (**i [enter]**). In [xxx](#ch.memory) we will look at how to decipher this information.

You can see all loaded modules and their sizes with **l [enter]** and the system version with **v [enter]**, while **k [enter]** will let you step through all processes and inspect them and kill them. Capital **D [enter]** will show you information about all the ETS tables in the system and lower case **d [enter]** will show you information about the distribution. That is basically just the node name.

If you have built your runtime with OPPROF or DEBUG you will be able to get even more information. We will look at how to do this in [CH-BuildingERTS]. The code for the break mode can be found in <filename>[OTP\_SOURCE]/erts/emulator/beam/break.c</filename>.

Note that going into break mode freezes the node. This is not something you want to do on a production system. But when debugging or profiling in a test system, this mode can help us find bugs and bottlenecks, as we will see later in this book.

## Chapter 21

---

# Tweaking the Runtime System



# Appendix A: Building the Erlang Runtime System

---

In this chapter we will look at different way to configure and build Erlang/OTP to suite your needs. We will use an Ubuntu Linux for most of the examples. If you are using a different OS you can find detailed instructions on how to build for that OS in the documentation in the source code (in HOWTO/INSTALL.md), or on the web [INSTALL.html](#) [[http://www.erlang.org/doc/installation\\_guide/INSTALL.html](http://www.erlang.org/doc/installation_guide/INSTALL.html)].

There are basically two ways to build the runtime system, the traditional way with autoconf, configure and make or with the help of [kerl](#) [<https://github.com/spawngrid/kerl>].

I recommend that you first give the traditional way a try, that way you will get a better understanding of what happens when you build and what settings you can change. Then go over to using kerl for your day to day job of managing configurations and builds.

## A.1. First Time Build

To get you started we will go though a step by step process of building the system from scratch and then we will look at how you can configure your system for different purposes.

This step by step guide assumes that you have a modern Ubuntu installation. We will look at how to build on OS X and Windows later in this chapter.

### A.1.1. Prerequisites

You will need a number of tools in order to fetch, unpack and build from source. The file Install.md lists some of the most important ones.

Given that we have a recent Ubuntu installation to start with many of the needed tools such as tar, make, perl and gcc should already be installed. But some tools like git, m4 and ncurses will probably need to be installed.

If you add a source URI to your apt configuration you will be able to use the build-dep command to get the needed sources to build erlang. You can do this by uncommenting the deb-src line for your distribution in /etc/apt/sources.list.

For the Yakkety Yak release you could add the line by:

```
> sudo cat "deb-src http://se.archive.ubuntu.com/ubuntu/ \
yakkety main restricted" >> /etc/apt/sources.list
```

Then the following commands will get almost all the tools you need:

```
> sudo apt-get install git autoconf m4
> sudo apt-get build-dep erlang
```

If you have a slightly older version of Ubuntu like Saucy and you want to build with wx support, you need to get the wx libraries:

```
> sudo apt-key adv --fetch-keys http://repos.codelite.org/CodeLite.asc
> sudo apt-add-repository 'deb http://repos.codelite.org/wx3.0/ubuntu/ saucy universe'
> sudo apt-get update
> sudo apt-get install libwxbase3.0-0-unofficial libwxbase3.0-dev libwxgtk3.0-0-unofficial \
libwxgtk3.0-dev wx3.0-headers wx-common libwxbase3.0-dbg libwxgtk3.0-dbg wx3.0-i18n \
wx3.0-examples wx3.0-doc
```

You might also want to create a directory where you keep the source code and also install your home built version without interfering with any pre built and system wide installations.

```
> cd
> mkdir otp
```

## A.2. Getting the source

There are two main ways of getting the source. You can download a tarball from [erlang.org](http://www.erlang.org) [<http://www.erlang.org/download.html>] or you can check out the source code directly from Github.

If you want to quickly download a stable version of the source try:

```
> cd ~/otp
> wget http://erlang.org/download/otp_src_19.1.tar.gz
> tar -xzf otp_src_19.1.tar.gz
> cd otp_src_19.1
> export ERL_TOP=`pwd`
```

or if you want to be able to easily update to the latest bleeding edge or you want to contribute fixes back to the community you can check out the source through git:

```
> cd ~/otp
> git clone https://github.com/erlang/otp.git source
> cd source
> export ERL_TOP=`pwd`
> ./otp_build autoconf
```

Now you are ready to build and install Erlang:

```
> export LANG=C
> ./configure --prefix=$HOME/otp/install
> make
> make install
> export PATH=$HOME/otp/install/bin/:$PATH
> export ROOTDIR=$HOME/otp/install/
```

## A.3. Building with Kerl

An easier way to build especially if you want to have several different builds available to experiment with is to build with Kerl.

# Appendix B: BEAM Instructions

Here we will go through most of the instructions in the BEAM generic instruction set in detail. In the next section we list all instructions with a brief explanation generated from the documentation in the code (see [lib/compiler/src/genop.tab](#)).

## B.1. Functions and Labels

### B.1.1. label Lbl

Instruction number 1 in the generic instruction set is not really an instruction at all. It is just a module local label giving a name, or actually a number to the current position in the code.

Each label potentially marks the beginning of a basic block since it is a potential destination of a jump.

### B.1.2. func\_info Module Function Arity

The code for each function starts with a `func_info` instruction. This instruction is used for generating a function clause error, and the execution of the code in the function actually starts at the label following the `func_info` instruction.

Imagine a function with a guard:

```
id(I) when is_integer(I) -> I.
```

The Beam code for this function might look like:

```
{function, id, 1, 4}.
{label,3}.
{func_info,{atom,test1},{atom,id},1}.
{label,4}.
{test,is_integer,{f,3},[{x,0}]}.
return.
```

Here the meta information `{function, id, 1, 4}` tells us that execution of the `id/1` function will start at label 4. At label 4 we do an `is_integer` on `x0` and if we fail we jump to label 3 (`f3`) which points to the `func_info` instruction, which will generate a *function clause* exception. Otherwise we just fall through and return the argument (`x0`).

 Function info instruction points to an `Export` record (defined in [erts/emulator/beam/export.h](#)) and located somewhere else in memory. The few dedicated words of memory inside that record are used by the tracing mechanism to place a special trace instruction which will trigger for each entry/return from the function by all processes.

## B.2. Test instructions

### B.2.1. Type tests

The type test instructions (`is_*` `Lbl Argument`) checks whether the argument is of the given type and if not jumps to the label `Lbl`. The beam disassembler wraps all these instructions in a `test` instruction. E.g.:

```
{test,is_integer,{f,3},[{x,0}]}.
```

The current type test instructions are `is_integer`, `is_float`, `is_number`, `is_atom`, `is_pid`, `is_reference`, `is_port`, `is_nil`, `is_binary`, `is_list`, `is_nonempty_list`, `is_function`, `is_function2`, `is_boolean`, `is_bitstr`, and `is_tuple`.

And then there is also one type test instruction of Arity 3: `test_arity Lbl Arg Arity`. This instruction tests that the arity of the argument (assumed to be a tuple) is of `Arity`. This instruction is usually preceded by an `is_tuple` instruction.

## B.2.2. Comparisons

The comparison instructions (`is_<op> Lbl Arg1 Arg2`) compares the two arguments according to the instructions and jumps to `Lbl` if the comparison fails.

The comparison instructions are: `is_lt`, `is_ge`, `is_eq`, `is_ne`, `is_eq_exact`, and `is_ne_exact`.

Remember that all Erlang terms are ordered so these instructions can compare any two terms. You can for example test if the atom `self` is less than the pid returned by `self()`. (It is.)

Note that for numbers the comparison is done on the Erlang type *number*, see [The Erlang Type System and Tags](#). That is, for a mixed float and integer comparison the number of lower precision is converted to the other type before comparison. For example on my system 1 and 1.0 compares as equal, as well as 9999999999999999 and 1.0e16. Comparing floating point numbers is always risk and best avoided, the result may vary depending on the underlying hardware.

If you want to make sure that the integer 1 and the floating point number 1.0 are compared different you can use `is_eq_exact` and `is_ne_exact`. This corresponds to the Erlang operators `=:=` and `=/=`.

## B.3. Function Calls

In this chapter we will summarize what the different call instructions does. For a thorough description of how function calls work see [Different Types of Calls, Linking and Hot Code Loading \(5p\)](#).

### B.3.1. call Arity Label

Does a call to the function of arity `Arity` in the same module at label `Label`. First count down the reductions and if needed do a context switch. Current code address after the call is saved into CP.

For all local calls the label is the second label of the function where the code starts. It is assumed that the preceding instruction at that label is `func_info` in order to get the MFA if a context switch is needed.

### B.3.2. call\_only Arity Label

Do a tail recursive call the function of arity `Arity` in the same module at label `Label`. First count down the reductions and if needed do a context switch. The CP is not updated with the return address.

### B.3.3. call\_last Arity Label Deallocate

Deallocate `Deallocate` words of stack, then do a tail recursive call to the function of arity `Arity` in the same module at label `Label`. First count down the reductions and if needed do a context switch. The CP is not updated with the return address.

### B.3.4. call\_ext Arity Destination

Does an external call to the function of arity `Arity` given by `Destination`. `Destination` in assembly is usually written as `{extfunc, Module, Function, Arity}`, this is then added to imports section of the module. First count down the reductions and if needed do a context switch. CP will be updated with the return address.

### B.3.5. call\_ext\_only Arity Destination

Does a tail recursive external call to the function of arity **Arity** given by Destination. Destination in assembly is usually written as **{extfunc, Module, Function, Arity}**. First count down the reductions and if needed do a context switch. The CP is not updated with the return address.

### B.3.6. call\_ext\_last Arity Destination Deallocate

Deallocate **Deallocate** words of stack, then do a tail recursive external call to the function of arity **Arity** given by Destination. Destination in assembly is usually written as **{extfunc, Module, Function, Arity}**. First count down the reductions and if needed do a context switch. The CP is not updated with the return address.

### B.3.7. bif0 Bif Reg, bif[1,2] Lbl Bif [Arg,...] Reg

Call the bif **Bif** with the given arguments, and store the result in **Reg**. If the bif fails, jump to **Lbl**. Zero arity bif cannot fail and thus **bif0** doesn't take a fail label.



Bif called by these instructions may not allocate on the heap nor trigger a garbage collection. Otherwise see: **gc\_bif**.

### B.3.8. gc\_bif[1-3] Lbl Live Bif [Arg, ...] Reg

Call the bif **Bif** with the given arguments, and store the result in **Reg**. If the bif fails, jump to **Lbl**. Arguments will be stored in **x(Live)**, **x(Live+1)** and **x(Live+2)**.



Because this instruction has argument **Live**, it gives us enough information to be able to trigger the garbage collection.

### B.3.9. call\_fun Arity

The instruction **call\_fun** assumes that the arguments are placed in the first **Arity** argument registers and that the fun (the pointer to the closure) is placed in the register following the last argument **x[Arity+1]**.

That is, for a zero arity call, the closure is placed in **x[0]**. For a arity 1 call **x[0]** contains the argument and **x[1]** contains the closure and so on.



Raises **badarity** if the arity doesn't match the function object. Raises **badfun** if a non-function is passed.

### B.3.10. apply Arity

Applies function call with **Arity** arguments stored in X registers. The module atom is stored in **x[Arity]** and the function atom is stored in **x[Arity+1]**. Module can also be represented by a tuple.

### B.3.11. apply\_last Arity Dealloc

Deallocates **Dealloc** elements on stack by popping CP, freeing the elements and pushing CP again. Then performs a tail-recursive call with **Arity** arguments stored in X registers, by jumping to the new location. The module and function atoms are stored in **x[Arity]** and **x[Arity+1]**. Module can also be represented by a tuple.

## B.4. Stack (and Heap) Management

The stack and the heap of an Erlang process on Beam share the same memory area see [Processes](#) and [The Memory Subsystem: Stacks, Heaps and Garbage Collection](#) for a full discussion. The stack grows toward lower addresses and the heap toward higher

addresses. Beam will do a garbage collection if more space than what is available is needed on either the stack or the heap.

### A leaf function

A leaf function is a function which doesn't call any other function.

### A non leaf function

A non leaf function is a function which may call another function.

These instructions are also used by non leaf functions for setting up and tearing down the stack frame for the current instruction. That is, on entry to the function the *continuation pointer* (CP) is saved on the stack, and on exit it is read back from the stack.

A function skeleton for a leaf function looks like this:

```
{function, Name, Arity, StartLabel}.
{label,L1}.
{func_info,{atom,Module},{atom,Name},Arity}.
{label,L2}.
...
return.
```

A function skeleton for a non leaf function looks like this:

```
{function, Name, Arity, StartLabel}.
{label,L1}.
{func_info,{atom,Module},{atom,Name},Arity}.
{label,L2}.
{allocate,Need,Live}.

...
call ...
...

{deallocate,Need}.
return.
```

#### B.4.1. allocate StackNeed Live

Save the continuation pointer (CP) and allocate space for **StackNeed** extra words on the stack. If during allocation we run out of memory, call the GC and then first **Live** x registers will form a part of the root set. E.g. if **Live** is 2 then GC will save registers X0 and X1, rest are unused and will be freed.

When allocating on the stack, the stack pointer (E) is decreased.

*Example 1. Allocate 1 0*

Before	After
E ->   xxx	xxx
xxx	???
	caller save slot
...    E ->   CP	...
...    HTOP ->	...
HTOP ->	xxx
xxx	

#### B.4.2. allocate\_heap StackNeed HeapNeed Live

Save the continuation pointer (CP) and allocate space for **StackNeed** extra words on the stack. Ensure that there also is space for **HeapNeed** words on the heap. If during allocation we run out of memory, call the GC with **Live** amount of X registers to preserve.



The heap pointer (HTOP) is not changed until the actual heap allocation takes place.

#### B.4.3. allocate\_zero StackNeed Live

This instruction works the same way as allocate, but it also clears out the allocated stack slots with **NIL**.

*Example 2. allocate\_zero 1 0*

Before	After
E ->   xxx	xxx
xxx	xxx
	NIL
	caller save slot
...    E ->   CP	...
...    HTOP ->	...
HTOP ->	xxx
xxx	

#### B.4.4. allocate\_heap\_zero StackNeed HeapNeed Live

The `allocate_heap_zero` instruction works as the `allocate_heap` instruction, but it also clears out the allocated stack slots with **NIL**.

#### B.4.5. test\_heap HeapNeed Live

The `test_heap` instruction ensures there is space for **HeapNeed** words on the heap. If during allocation we run out of memory, call the GC with **Live** amount of X registers to preserve.

#### B.4.6. init N

The `init` instruction clears N stack words above the CP pointer by writing **NIL** to them.

### B.4.7. deallocate N

The `deallocate` instruction is the opposite of the `allocate`. It restores the CP (continuation pointer) and deallocates `N+1` stack words.

### B.4.8. return

The return instructions jumps to the address in the continuation pointer (CP). The value of CP is set to `0` in C.

### B.4.9. trim N Remaining

Pops the CP into a temporary variable, frees `N` words of stack, and places the CP back onto the top of the stack. (The argument `Remaining` is to the best of my knowledge unused.)

*Example 3. Trim 2*

Before	After
???	???
xxx	E ->   CP
xxx	...
E ->   CP	...
	...
...	...
HTOP ->	HTOP ->
xxx	xxx

## B.5. Moving, extracting, modifying data

### B.5.1. move Source Destination

Moves the value of the source `Source` (this can be a literal or a register) to the destination register `Destination`.

### B.5.2. get\_list Source Head Tail

This is a deconstruct operation for a list cell. Get the head and tail (or car and cdr) parts of a list (a cons cell), specified by `Source` and place them into the registers `Head` and `Tail`.

### B.5.3. get\_tuple\_element Source Element Destination

This is an array indexed read operation. Get element with position `Element` from the `Source` tuple and place it into the `Destination` register.

### B.5.4. set\_tuple\_element NewElement Tuple Position

This is a destructive array indexed update operation. Update the element of the `Tuple` at `Position` with the new `NewElement`.

## B.6. Building terms.

## B.6.1. put\_list Head Tail Destination

Constructs a new list (cons) cell on the heap (2 words) and places its address into the **Destination** register. First element of list cell is set to the value of **Head**, second element is set to the value of **Tail**.

## B.6.2. put\_tuple Size Destination

Constructs an empty tuple on the heap (**Size+1** words) and places its address into the **Destination** register. No elements are set at this moment. **Put\_tuple** instruction is always followed by multiple **put** instructions which destructively set its elements one by one.

## B.6.3. put Value

Places destructively a **Value** into the next element of a tuple, which was created by a preceding **put\_tuple** instruction. Write address is maintained and incremented internally by the VM. Multiple **put** instructions are used to set contents for any new tuple.

## B.6.4. make\_fun2 LambdaIndex

Creates a function object defined by an index in the Lambda table of the module. A lambda table defines the entry point (a label or export entry), arity and how many frozen variables to take. Frozen variable values are copied from the current execution context (X registers) and stored into the function object.

# B.7. Binary Syntax

## B.7.1. bs\_put\_integer/5

TODO

## B.7.2. bs\_put\_binary/5

TODO

## B.7.3. bs\_put\_float/5

TODO

## B.7.4. bs\_put\_string/2

TODO

## B.7.5. bs\_init2/6

TODO

## B.7.6. bs\_add/5

TODO

## B.7.7. bs\_start\_match2/5

TODO

## **B.7.8. bs\_get\_integer2/7**

TODO

## **B.7.9. bs\_get\_float2/7**

TODO

## **B.7.10. bs\_get\_binary2/7**

TODO

## **B.7.11. bs\_skip\_bits2/5**

TODO

## **B.7.12. bs\_test\_tail2/3**

TODO

## **B.7.13. bs\_save2/2**

TODO

## **B.7.14. bs\_restore2/2**

TODO

## **B.7.15. bs\_context\_to\_binary/1**

TODO

## **B.7.16. bs\_test\_unit/3**

TODO

## **B.7.17. bs\_match\_string/4**

TODO

## **B.7.18. bs\_init\_writable/0**

TODO

## **B.7.19. bs\_append/8**

TODO

## **B.7.20. bs\_private\_append/6**

TODO

### B.7.21. `bs_init_bits/6`

TODO

### B.7.22. `bs_get_utf8/5`

TODO

### B.7.23. `bs_skip_utf8/4`

TODO

### B.7.24. `bs_get_utf16/5`

TODO

### B.7.25. `bs_skip_utf16/4`

TODO

### B.7.26. `bs_get_utf32/5`

TODO

### B.7.27. `bs_skip_utf32/4`

TODO

### B.7.28. `bs_utf8_size/3`

TODO

### B.7.29. `bs_put_utf8/3`

TODO

### B.7.30. `bs_utf16_size/3`

TODO

### B.7.31. `bs_put_utf16/3`

TODO

### B.7.32. `bs_put_utf32/3`

TODO

## B.8. Floating Point Arithmetic

### B.8.1. `fclearerror/0`

TODO

### **B.8.2. fcheckerror/1**

TODO

### **B.8.3. fmove/2**

TODO

### **B.8.4. fconv/2**

TODO

### **B.8.5. fadd/4**

TODO

### **B.8.6. fsub/4**

TODO

### **B.8.7. fmul/4**

TODO

### **B.8.8. fdiv/4**

TODO

### **B.8.9. fnegate/3**

TODO

## **B.9. Pattern Matching**

### **B.9.1. select\_val**

TODO

### **B.9.2. select\_arity\_val**

TODO

### **B.9.3. jump**

TODO

## **B.10. Exception handling**

### **B.10.1. catch/2**

TODO

## B.10.2. catch\_end/1

TODO

## B.10.3. badmatch/1

TODO

## B.10.4. if\_end/0

TODO

## B.10.5. case\_end/1

TODO

# B.11. Meta instructions

## B.11.1. on\_load

TODO

## B.11.2. line

TODO

# B.12. Generic Instructions

Name	Arity	Op Code	Spec	Documentation
allocate	2	12	<b>allocate</b> <i>StackNeed, Live</i>	Allocate space for StackNeed words on the stack. If a GC is needed during allocation there are Live number of live X registers. Also save the continuation pointer (CP) on the stack.
allocate_heap	3	13	<b>allocate_heap</b> <i>StackNeed, HeapNeed, Live</i>	Allocate space for StackNeed words on the stack and ensure there is space for HeapNeed words on the heap. If a GC is needed save Live number of X registers. Also save the continuation pointer (CP) on the stack.

Name	Arity	Op Code	Spec	Documentation
allocate_heap_zero	3	15	<b>allocate_heap_zero</b> <i>StackNeed, HeapNeed, Live</i>	Allocate space for StackNeed words on the stack and HeapNeed words on the heap. If a GC is needed during allocation there are Live number of live X registers. Clear the new stack words. (By writing NIL.) Also save the continuation pointer (CP) on the stack.
allocate_zero	2	14	<b>allocate_zero</b> <i>StackNeed, Live</i>	Allocate space for StackNeed words on the stack. If a GC is needed during allocation there are Live number of live X registers. Clear the new stack words. (By writing NIL.) Also save the continuation pointer (CP) on the stack.
apply	1	112		
apply_last	2	113		
badmatch	1	72		
bif0	2	9	<b>bif0</b> <i>Bif, Reg</i>	Call the bif Bif and store the result in Reg.
bif1	4	10	<b>bif1</b> <i>Lbl, Bif, Arg, Reg</i>	Call the bif Bif with the argument Arg, and store the result in Reg. On failure jump to Lbl.
bif2	5	11	<b>bif2</b> <i>Lbl, Bif, Arg1, Arg2, Reg</i>	Call the bif Bif with the arguments Arg1 and Arg2, and store the result in Reg. On failure jump to Lbl.
bs_add	5	111		
bs_append	8	134		
<del>bs_bits_to_bytes</del>	3	(110)	<b>DEPRECATED</b>	
<del>bs_bits_to_bytes2</del>	2	(127)	<b>DEPRECATED</b>	
bs_context_to_binary	1	130		
<del>bs_final</del>	2	(88)	<b>DEPRECATED</b>	
<del>bs_final2</del>	2	(126)	<b>DEPRECATED</b>	
<del>bs_get_binary</del>	5	(82)	<b>DEPRECATED</b>	
bs_get_binary2	7	119		

Name	Arity	Op Code	Spec	Documentation
<code>bs_get_float</code>	5	(81)	<b>DEPRECATED</b>	
<code>bs_get_float2</code>	7	118		
<code>bs_get_integer</code>	5	(80)	<b>DEPRECATED</b>	
<code>bs_get_integer2</code>	7	117		
<code>bs_get_position</code>	3	167	<code>bs_get_position Ctx, Dst, Live</code>	Sets Dst to the current position of Ctx
<code>bs_get_tail</code>	3	165	<code>bs_get_tail Ctx, Dst, Live</code>	Sets Dst to the tail of Ctx at the current position
<code>bs_get_utf16</code>	5	140		
<code>bs_get_utf32</code>	5	142		
<code>bs_get_utf8</code>	5	138		
<code>bs_init</code>	2	(87)	<b>DEPRECATED</b>	
<code>bs_init2</code>	6	109		
<code>bs_init_bits</code>	6	137		
<code>bs_init_writable</code>	0	133		
<code>bs_match_string</code>	4	132		
<code>bs_need_buf</code>	1	(93)	<b>DEPRECATED</b>	
<code>bs_private_append</code>	6	135		
<code>bs_put_binary</code>	5	90		
<code>bs_put_float</code>	5	91		
<code>bs_put_integer</code>	5	89		
<code>bs_put_string</code>	2	92		
<code>bs_put_utf16</code>	3	147		
<code>bs_put_utf32</code>	3	148		
<code>bs_put_utf8</code>	3	145		
<code>bs_restore</code>	1	(86)	<b>DEPRECATED</b>	
<code>bs_restore2</code>	2	123		
<code>bs_save</code>	1	(85)	<b>DEPRECATED</b>	
<code>bs_save2</code>	2	122		
<code>bs_set_position</code>	2	168	<code>bs_set_positon Ctx, Pos</code>	Sets the current position of Ctx to Pos
<code>bs_skip_bits</code>	4	(83)	<b>DEPRECATED</b>	
<code>bs_skip_bits2</code>	5	120		
<code>bs_skip_utf16</code>	4	141		

Name	Arity	Op Code	Spec	Documentation
bs_skip_utf32	4	143		
bs_skip_utf8	4	139		
<b>bs_start_match</b>	2	(79)	<b>DEPRECATED</b>	
bs_start_match2	5	116		
bs_start_match3	4	166	<b>bs_start_match3</b> Fail, Bin, Live, Dst	Starts a binary match sequence
bs_start_match4	4	170	<b>bs_start_match4</b> Fail, Bin, Live, Dst	As bs_start_match3, but the fail label can be 'no_fail' when we know it will never fail at runtime, or 'resume' when we know the input is a match context.
<b>bs_test_tail</b>	2	(84)	<b>DEPRECATED</b>	
bs_test_tail2	3	121		
bs_test_unit	3	131		
bs_utf16_size	3	146		
bs_utf8_size	3	144		
build_stacktrace	0	160	<b>build_stacktrace</b>	Given the raw stacktrace in x(0), build a cooked stacktrace suitable for human consumption. Store it in x(0). Destroys all other registers. Do a garbage collection if necessary to allocate space on the heap for the result.
call	2	4	<b>call</b> Arity, Label	Call the function at Label. Save the next instruction as the return address in the CP register.
call_ext	2	7	<b>call_ext</b> Arity, Destination	Call the function of arity Arity pointed to by Destination. Save the next instruction as the return address in the CP register.
call_ext_last	3	8	<b>call_ext_last</b> Arity, Destination, Deallocate	Deallocate and do a tail call to function of arity Arity pointed to by Destination. Do not update the CP register. Deallocate words from the stack before the call.

Name	Arity	Op Code	Spec	Documentation
call_ext_only	2	78	<b>call_ext_only</b> Arity, <i>Label</i>	Do a tail recursive call to the function at Label. Do not update the CP register.
call_fun	1	75	<b>call_fun</b> Arity	Call a fun of arity Arity. Assume arguments in registers x(0) to x(Arity-1) and that the fun is in x(Arity). Save the next instruction as the return address in the CP register.
call_last	3	5	<b>call_last</b> Arity, <i>Label</i> , Deallocate	Deallocate and do a tail recursive call to the function at Label. Do not update the CP register. Before the call deallocate Deallocate words of stack.
call_only	2	6	<b>call_only</b> Arity, <i>Label</i>	Do a tail recursive call to the function at Label. Do not update the CP register.
case_end	1	74		
catch	2	62		
catch_end	1	63		
deallocate	1	18	<b>deallocate</b> N	Restore the continuation pointer (CP) from the stack and deallocate N+1 words from the stack (the + 1 is for the CP).
fadd	4	98		
fcheckerror	1	95		
fclearerror	0	94		
fconv	2	97		
fdiv	4	101		
fmove	2	96		
fmul	4	100		
fnegate	3	102		
fsub	4	99		
func_info	3	2	<b>func_info</b> M, F, A	Define a function M:F/A

Name	Arity	Op Code	Spec	Documentation
gc_bif1	5	124	<b>gc_bif1</b> <i>Lbl, Live, Bif, Arg, Reg</i>	Call the bif Bif with the argument Arg, and store the result in Reg. On failure jump to Lbl. Do a garbage collection if necessary to allocate space on the heap for the result (saving Live number of X registers).
gc_bif2	6	125	<b>gc_bif2</b> <i>Lbl, Live, Bif, Arg1, Arg2, Reg</i>	Call the bif Bif with the arguments Arg1 and Arg2, and store the result in Reg. On failure jump to Lbl. Do a garbage collection if necessary to allocate space on the heap for the result (saving Live number of X registers).
gc_bif3	7	152	<b>gc_bif3</b> <i>Lbl, Live, Bif, Arg1, Arg2, Arg3, Reg</i>	Call the bif Bif with the arguments Arg1, Arg2 and Arg3, and store the result in Reg. On failure jump to Lbl. Do a garbage collection if necessary to allocate space on the heap for the result (saving Live number of X registers).
get_hd	2	162	<b>get_hd</b> <i>Source, Head</i>	Get the head (or car) part of a list (a cons cell) from Source and put it into the register Head.
get_list	3	65	<b>get_list</b> <i>Source, Head, Tail</i>	Get the head and tail (or car and cdr) parts of a list (a cons cell) from Source and put them into the registers Head and Tail.
get_map_elements	3	158		
get_tl	2	163	<b>get_tl</b> <i>Source, Tail</i>	Get the tail (or cdr) part of a list (a cons cell) from Source and put it into the register Tail.
get_tuple_element	3	66	<b>get_tuple_element</b> <i>Source, Element, Destination</i>	Get element number Element from the tuple in Source and put it in the destination register Destination.
has_map_fields	3	157		

Name	Arity	Op Code	Spec	Documentation
if_end	0	73		
init	1	17	init <i>N</i>	Clear the Nth stack word. (By writing NIL.)
<del>int_band</del>	4	(33)	<b>DEPRECATED</b>	
<del>int_bnot</del>	3	(38)	<b>DEPRECATED</b>	
<del>int_bor</del>	4	(34)	<b>DEPRECATED</b>	
<del>int_bsl</del>	4	(36)	<b>DEPRECATED</b>	
<del>int_bsr</del>	4	(37)	<b>DEPRECATED</b>	
<del>int_bxor</del>	4	(35)	<b>DEPRECATED</b>	
int_code_end	0	3		
<del>int_div</del>	4	(31)	<b>DEPRECATED</b>	
<del>int_rem</del>	4	(32)	<b>DEPRECATED</b>	
is_atom	2	48	is_atom <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not an atom.
is_binary	2	53	is_binary <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a binary.
is_bitstr	2	129	is_bitstr <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a bit string.
is_boolean	2	114	is_boolean <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a Boolean.
<del>is_constant</del>	2	(54)	<b>DEPRECATED</b>	
is_eq	3	41	is_eq <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is not (numerically) equal to Arg2.
is_eq_exact	3	43	is_eq_exact <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is not exactly equal to Arg2.
is_float	2	46	is_float <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a float.
is_function	2	77	is_function <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a function (i.e. fun or closure).

Name	Arity	Op Code	Spec	Documentation
is_function2	3	115	<b>is_function2</b> <i>Lbl, Arg1, Arity</i>	Test the type of Arg1 and jump to Lbl if it is not a function of arity Arity.
is_ge	3	40	<b>is_ge</b> <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is less than Arg2.
is_integer	2	45	<b>is_integer</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not an integer.
is_list	2	55	<b>is_list</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a cons or nil.
is_lt	3	39	<b>is_lt</b> <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is not less than Arg2.
is_map	2	156		
is_ne	3	42	<b>is_ne</b> <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is (numerically) equal to Arg2.
is_ne_exact	3	44	<b>is_ne_exact</b> <i>Lbl, Arg1, Arg2</i>	Compare two terms and jump to Lbl if Arg1 is exactly equal to Arg2.
is_nil	2	52	<b>is_nil</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not nil.
is_nonempty_list	2	56	<b>is_nonempty_list</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a cons.
is_number	2	47	<b>is_number</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a number.
is_pid	2	49	<b>is_pid</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a pid.
is_port	2	51	<b>is_port</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a port.
is_reference	2	50	<b>is_reference</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a reference.

Name	Arity	Op Code	Spec	Documentation
is_tagged_tuple	4	159	<b>is_tagged_tuple</b> <i>Lbl, Reg, N, Atom</i>	Test the type of Reg and jumps to Lbl if it is not a tuple. Test the arity of Reg and jumps to Lbl if it is not N. Test the first element of the tuple and jumps to Lbl if it is not Atom.
is_tuple	2	57	<b>is_tuple</b> <i>Lbl, Arg1</i>	Test the type of Arg1 and jump to Lbl if it is not a tuple.
jump	1	61	<b>jump</b> <i>Label</i>	Jump to Label.
label	1	1	<b>label</b> <i>Lbl</i>	Specify a module local label. Label gives this code address a name (Lbl) and marks the start of a basic block.
line	1	153		
loop_rec	2	23	<b>loop_rec</b> <i>Label, Source</i>	Loop over the message queue, if it is empty jump to Label.
loop_rec_end	1	24	<b>loop_rec_end</b> <i>Label</i>	Advance the save pointer to the next message and jump back to Label.
<del>m_div</del>	4	(30)	<b>DEPRECATED</b>	
<del>m_minus</del>	4	(28)	<b>DEPRECATED</b>	
<del>m_plus</del>	4	(27)	<b>DEPRECATED</b>	
<del>m_times</del>	4	(29)	<b>DEPRECATED</b>	
<del>make_fun</del>	3	(76)	<b>DEPRECATED</b>	
make_fun2	1	103		
move	2	64	<b>move</b> <i>Source, Destination</i>	Move the source Source (a literal or a register) to the destination register Destination.
on_load	0	149		
put	1	71		
put_list	3	69		
<del>put_literal</del>	2	(128)	<b>DEPRECATED</b>	
put_map_assoc	5	154		
put_map_exact	5	155		
<del>put_string</del>	3	(68)	<b>DEPRECATED</b>	

Name	Arity	Op Code	Spec	Documentation
put_tuple	2	70		
put_tuple2	2	164	<b>put_tuple2</b> <i>Destination, Elements</i>	Build a tuple with the elements in the list Elements and put it into register Destination.
raise	2	108		
raw_raise	0	161	<b>raw_raise</b>	This instruction works like the erlang:raise/3 BIF, except that the stacktrace in x(2) must be a raw stacktrace. x(0) is the class of the exception (error, exit, or throw), x(1) is the exception term, and x(2) is the raw stackframe. If x(0) is not a valid class, the instruction will not throw an exception, but store the atom 'badarg' in x(0) and execute the next instruction.
recv_mark	1	150	<b>recv_mark</b> <i>Label</i>	Save the end of the message queue and the address of the label Label so that a recv_set instruction can start scanning the inbox from this position.
recv_set	1	151	<b>recv_set</b> <i>Label</i>	Check that the saved mark points to Label and set the save pointer in the message queue to the last position of the message queue saved by the recv_mark instruction.
remove_message	0	21	<b>remove_message</b>	Unlink the current message from the message queue. Remove any timeout.
return	0	19	<b>return</b>	Return to the address in the continuation pointer (CP).
select_tuple_arity	3	60	<b>select_tuple_arity</b> <i>Tuple, FailLabel, Destinations</i>	Check the arity of the tuple Tuple and jump to the corresponding destination label, if no arity matches, jump to FailLabel.

Name	Arity	Op Code	Spec	Documentation
select_val	3	59	<b>select_val</b> Arg, <i>FailLabel, Destinations</i>	Jump to the destination label corresponding to Arg in the Destinations list, if no arity matches, jump to FailLabel.
send	0	20	<b>send</b>	Send argument in x(1) as a message to the destination process in x(0). The message in x(1) ends up as the result of the send in x(0).
set_tuple_element	3	67	<b>set_tuple_element</b> <i>NewElement, Tuple, Position</i>	Update the element at position Position of the tuple Tuple with the new element NewElement.
swap	2	169	<b>swap</b> Register1, Register2	Swaps the contents of two registers.
test_arity	3	58	<b>test_arity</b> Lbl, Arg1, Arity	Test the arity of (the tuple in) Arg1 and jump to Lbl if it is not equal to Arity.
test_heap	2	16	<b>test_heap</b> HeapNeed, Live	Ensure there is space for HeapNeed words on the heap. If a GC is needed save Live number of X registers.
timeout	0	22	<b>timeout</b>	Reset the save point of the mailbox and clear the timeout flag.
trim	2	136	<b>trim</b> N, Remaining	Reduce the stack usage by N words, keeping the CP on the top of the stack.
try	2	104		
try_case	1	106		
try_case_end	1	107		
try_end	1	105		
wait	1	25	<b>wait</b> Label	Suspend the processes and set the entry point to the beginning of the receive loop at Label.
wait_timeout	2	26	<b>wait_timeout</b> Label, Time	Sets up a timeout of Time milliseconds and saves the address of the following instruction as the entry point if the timeout triggers.

## B.13. Specific Instructions

Argument types

Typ	Explanation
e	
a	An immediate atom value, e.g. 'foo'
c	An immediate constant value (atom, nil, small int) // Pid?
d	Either a register or a stack slot
e	A reference to an export table entry
f	A label, i.e. a code address
I	An integer e.g. 42
j	An optional code label
l	A floating-point register
P	A positive (unsigned) integer literal
r	A register R0 ( <code>x[0]</code> )
s	Either a literal, a register or a stack slot
t	A term, e.g. <code>[{foo, bar}]</code>
x	A register, e.g. 5 for <code>{x, 5}</code>
y	A stack slot, e.g. 1 for <code>{y, 1}</code>

### B.13.1. List of all BEAM Instructions

Instruction	Arguments	Explanation
allocate	t t	Allocate some words on stack
allocate_heap	t I t	Allocate some words on the heap
allocate_heap_zero	t I t	Allocate some heap and set the words to NIL
allocate_init	t I y	
allocate_zero	t t	Allocate some stack and set the words to 0?
apply	I	Apply args in <code>x[0..Arity-1]</code> to module in <code>x[Arity]</code> and function in <code>x[Arity+1]</code>
apply_last	I P	Same as <code>apply</code> but does not save the CP and <code>deallocates</code> P words
badarg	j	Create a <code>badarg</code> error
badmatch	rxy	Create a <code>badmatch</code> error
bif1	f b s d	Calls a bif with 1 argument, on fail jumps to <code>f</code>
bif1_body	b s d	
bs_context_to_binary	rxy	
bs_put_string	I I	

Instruction	Arguments	Explanation
bs_test_tail_imm2	f rx I	
bs_test_unit	f rx I	
bs_test_unit8	f rx	
bs_test_zero_tail2	f rx	
call_bif0	e	
call_bif1	e	
call_bif2	e	
call_bif3	e	
case_end	rxy	Create a <b>case_clause</b> error
catch	y f	
catch_end	y	
deallocate	I	Free some words from stack and pop CP
deallocate_return	Q	Combines <b>deallocate</b> and <b>return</b>
extract_next_element	xy	
extract_next_element_2	xy	
extract_next_element_3	xy	
fclearerror		
fconv	d l	
fmove	qdl ld	
get_list	rxy rxy rxy	Deconstruct a list cell into the head and the tail
i_apply		Call the code for function <b>x0:x1</b> with args <b>x2</b> saving the CP
i_apply_fun		Call the code for function object <b>x0</b> with args <b>x1</b> saving the CP
i_apply_fun_last	P	Jump to the code for function object <b>x0</b> with args <b>x1</b> , restoring the CP and deallocating <b>P</b> stack cells
i_apply_fun_only		Jump to the code for function object <b>x0</b> with args <b>x1</b>
i_apply_last	P	Jump to the code for function <b>x0:x1</b> with args <b>x2</b>
i_apply_only		Jump to the code for function <b>x0:x1</b> with args <b>x2</b>
i_band	j I d	
i_bif2	f b d	
i_bif2_body	b d	
i_bor	j I d	
i_bs_add	j I d	

Instruction	Arguments	Explanation
i_bs_append	j I I I d	
i_bs_get_binary2	f rx I s I d	
i_bs_get_binary_all2	f rx I I d	
i_bs_get_binary_all_reuse	rx f I	
i_bs_get_binary_imm2	f rx I I I d	
i_bs_get_float2	f rx I s I d	
i_bs_get_integer	f I I d	
i_bs_get_integer_16	rx f d	
i_bs_get_integer_32	rx f I d	
i_bs_get_integer_8	rx f d	
i_bs_get_integer_imm	rx I I f I d	
i_bs_get_integer_small_imm	rx I f I d	
i_bs_get_utf16	rx f I d	
i_bs_get_utf8	rx f d	
i_bs_init	I I d	
i_bs_init_bits	I I d	
i_bs_init_bits_fail	rxy j I d	
i_bs_init_bits_fail_heap	I j I d	
i_bs_init_bits_heap	I I I d	
i_bs_init_fail	rxy j I d	
i_bs_init_fail_heap	I j I d	
i_bs_init_heap	I I I d	
i_bs_init_heap_bin	I I d	
i_bs_init_heap_bin_heap	I I I d	
i_bs_init_writable		
i_bs_match_string	rx f I I	
i_bs_private_append	j I d	
i_bs_put_utf16	j I s	

Instruction	Arguments	Explanation
i_bs_put_utf8	j s	
i_bs_restore2	rx I	
i_bs_save2	rx I	
i_bs_skip_bits2	f rx rxy I	
i_bs_skip_bits2_imm2	f rx I	
i_bs_skip_bits_all2	f rx I	
i_bs_start_match2	rxy f II d	
i_bs_utf16_size	s d	
i_bs_utf8_size	s d	
i_bs_validate_unicode	j s	
i_bs_validate_unicode_retract	j	
i_bsl	j I d	
i_bsr	j I d	
i_bxor	j I d	
i_call	f	
i_call_ext	e	
i_call_ext_last	e P	
i_call_ext_only	e	
i_call_fun	I	
i_call_fun_last	I P	
i_call_last	f P	
i_call_only	f	
i_element	rxy j s d	
i_fadd	111	
i_fast_element	rxy j I d	
i_fcheckerror		
i_fdiv	111	
i_fetch	s s	
i_fmul	111	
i_fnegate	111	
i_fsub	111	

Instruction	Arguments	Explanation
i_func_info	I a a I	Create a <code>function_clause</code> error
i_gc_bif1	j I s I d	
i_gc_bif2	j I I d	
i_gc_bif3	j I s I d	
i_get	s d	
i_get_tuple_element	rxy P rxy	
i_hibernate		
i_increment	rxy I I d	
i_int_bnot	j s I d	
i_int_div	j I d	
i_is_eq	f	
i_is_eq_exact	f	
i_is_eq_exact_immed	f rxy c	
i_is_eq_exact_literal	f rxy c	
i_is_ge	f	
i_is_lt	f	
i_is_ne	f	
i_is_ne_exact	f	
i_is_ne_exact_immed	f rxy c	
i_is_ne_exact_literal	f rxy c	
i_jump_on_val	rxy f I I	
i_jump_on_val_zero	rxy f I	
i_loop_rec	f r	
i_m_div	j I d	
i_make_fun	I t	
i_minus	j I d	
i_move_call	c r f	
i_move_call_ext	c r e	
i_move_call_ext_last	e P c r	
i_move_call_ext_only	e c r	
i_move_call_last	f P c r	
i_move_call_only	f c r	
i_new_bs_put_binary	j s I s	

Instruction	Arguments	Explanation
i_new_bs_put_binary_all	j s I	
i_new_bs_put_binary_imm	j I s	
i_new_bs_put_float	j s I s	
i_new_bs_put_float_imm	j I I s	
i_new_bs_put_integer	j s I s	
i_new_bs_put_integer_imm	j I I s	
i_plus	j I d	
i_put_tuple	rxy I	Create tuple of arity <b>I</b> and place result in <b>rxy</b> , elements follow as <b>put</b> instructions
i_recv_set	f	
i_rem	j I d	
i_select_tuple_arity	r f I	
i_select_tuple_arity	x f I	
i_select_tuple_arity	y f I	
i_select_tuple_arity2	r f A f A f	
i_select_tuple_arity2	x f A f A f	
i_select_tuple_arity2	y f A f A f	
i_select_val	r f I	Compare value to a list of pairs <b>{Value, Label}</b> and jump when a match is found, otherwise jump to <b>f</b>
i_select_val	x f I	Same as above but for x register
i_select_val	y f I	Same as above but for y register
i_select_val2	r f c f c f	Compare value to two pairs <b>{c1, f1}</b> , or <b>{c2, f2}</b> and jump, on fail jump to <b>f</b>
i_select_val2	x f c f c f	Same as above but for x register
i_select_val2	y f c f c f	Same as above but for y register
i_times	j I d	
i_trim	I	Cut stack by <b>I</b> elements, preserving CP on top
i_wait_error		
i_wait_error_locked		
i_wait_timeout	f I	

Instruction	Arguments	Explanation
i_wait_timeout	f s	
i_wait_timeout_locke d	f I	
i_wait_timeout_locke d	f s	
if_end		Create an <b>if_clause</b> error
init	y	Set a word on stack to NIL []
init2	y y	Set two words on stack to NIL []
init3	y y y	Set three words on stack to NIL []
int_code_end		End of the program (same as return with no stack)
is_atom	f rxy	Check whether a value is an atom and jump otherwise
is_bitstring	f rxy	Check whether a value is a bit string and jump otherwise
is_boolean	f rxy	Check whether a value is atom 'true' or 'false' and jump otherwise
is_float	f rxy	Check whether a value is a floating point number and jump otherwise
is_function	f rxy	Check whether a value is a function and jump otherwise
is_function2	f s s	Check whether a value is a function and jump otherwise
is_integer	f rxy	Check whether a value is a big or small integer and jump otherwise
is_integer_allocate	f rx II	
is_list	f rxy	Check whether a value is a list or NIL and jump otherwise
is_nil	f rxy	Check whether a value is an empty list [] and jump otherwise
is_nonempty_list	f rxy	Check whether a value is a nonempty list (cons pointer) and jump otherwise
is_nonempty_list_all ocate	f rx It	
is_nonempty_list_te st_heap	f I t	
is_number	f rxy	Check whether a value is a big or small integer or a float and jump otherwise
is_pid	f rxy	Check whether a value is a pid and jump otherwise
is_port	f rxy	Check whether a value is a port and jump otherwise
is_reference	f rxy	Check whether a value is a reference and jump otherwise
is_tuple	f rxy	Check whether a value is a tuple and jump otherwise
is_tuple_of_arity	f rxy A	Check whether a value is a tuple of arity A and jump otherwise
jump	f	Jump to location (label) f
label	L	Marks a location in code, removed at the load time
line	I	Marks a location in source file, removed at the load time
loop_rec_end	f	Advances receive pointer in the process and jumps to the <b>loop_rec</b> instruction

Instruction	Arguments	Explanation
move	r sync r xy	Moves a value or a register into another register
move2	x x x x	Move a pair of values to a pair of destinations
move2	x y x y	Move a pair of values to a pair of destinations
move2	y x y x	Move a pair of values to a pair of destinations
move_call	xy r f	
move_call_last	xy r f Q	
move_call_only	x r f	
move_deallocate_return	xy cn r Q	
move_jump	f nc xy	
move_return	xcn r	
move_x1	c	Store value in $x_1$
move_x2	c	Store value in $x_2$
node	r xy	Get $\text{r } xy$ to the atom, current node name
put	r xy	Sequence of these is placed after $i\_put\_tuple$ and is used to initialize tuple elements (starting from 0)
put_list	s s d	Construct a list cell from a head and a tail and the cons pointer is placed into destination $d$
raise	s s	Raise an exception of given type, the exception type has to be extracted from the second stacktrace argument due to legacy/compatibility reasons.
recv_mark	f	Mark a known restart position for messages retrieval (reference optimization)
remove_message		Removes current message from the process inbox (was received)
return		Jump to the address in CP, set CP to 0
self	r xy	Set $\text{r } xy$ to the pid of the current process
send		Send message $x_1$ to the inbox of process $x_0$ , there is no error if process did not exist
set_tuple_element	s d P	Destructively update a tuple element by index
system_limit	j	
test arity	f r xy A	Check whether function object (closure or export) in $\text{r } xy$ has arity $A$ and jump to $f$ otherwise
test_heap	I t	Check the heap space availability
test_heap_1_put_list	I y	
timeout		Sets up a timer and yields the execution of the process waiting for an incoming message, or a timer event whichever comes first
timeout_locked		
try	y f	Writes a special catch value to stack cell $y$ which marks an active try block, the VM will jump to the label $f$ if an exception happens. Code which runs after this becomes guarded against exceptions

<b>Instruction</b>	<b>Arguments</b>	<b>Explanation</b>
try_case	y	Similar to <code>try_end</code> marks an end of the guarded section, clears the catch value on stack and begins the code section of exception matching
try_case_end	s	
try_end	y	Clears the catch value from the stack cell <code>y</code> marking an end of the guarded section
wait	f	Schedules the process out waiting for an incoming message (yields)
wait_locked	f	
wait_unlocked	f	

## Appendix C: Full Code Listings

```

-module(beanfile).
-export([read/1]).

read(Filename) ->
{ok, File} = file:read_file(Filename),
<<"FOR1",
  Size:32/integer,
  "BEAM",
  Chunks/binary>> = File,
{Size, parse_chunks(read_chunks(Chunks, []), [])}.

read_chunks(<<N,A,M,E, Size:32/integer, Tail/binary>>, Acc) ->
% Align each chunk on even 4 bytes
ChunkLength = align_by_four(Size),
<<Chunk:ChunkLength/binary, Rest/binary>> = Tail,
read_chunks(Rest, [{[N,A,M,E], Size, Chunk}|Acc]);
read_chunks(<<>>, Acc) -> lists:reverse(Acc).

align_by_four(N) -> (4 * ((N+3) div 4)).

parse_chunks([{{"Atom", _Size, <<_Numberofatoms:32/integer, Atoms/binary>>} | Rest], Acc) ->
  parse_chunks(Rest, [{atoms,parse_atoms(Atoms)}|Acc]);
parse_chunks([{{"ExpT", _Size,
  <<_Numberofentries:32/integer, Exports/binary>>}
  | Rest}, Acc] ->
  parse_chunks(Rest, [{exports,parse_table(Exports)}|Acc]);
parse_chunks([{{"ImpT", _Size,
  <<_Numberofentries:32/integer, Imports/binary>>}
  | Rest}, Acc] ->
  parse_chunks(Rest, [{imports,parse_table(Imports)}|Acc]);
parse_chunks([{{"Code", Size, <<SubSize:32/integer, Chunk/binary>>} | Rest], Acc) ->
  <<Info:SubSize/binary, Code/binary>> = Chunk,
  OpcodeSize = Size - SubSize - 8, %% 8 is size of CunkSize & SubSize
  <<OpCodes:OpcodeSize/binary, _Align/binary>> = Code,
  parse_chunks(Rest, [{code,parse_code_info(Info), OpCodes}|Acc]);
parse_chunks([{{"StrT", _Size, <<Strings/binary>>} | Rest], Acc) ->
  parse_chunks(Rest, [{strings,binary_to_list(Strings)}|Acc]);
parse_chunks([{{"Attr", Size, Chunk} | Rest}, Acc] ->
  <<Bin:Size/binary, _Pad/binary>> = Chunk,
  Attribs = binary_to_term(Bin),
  parse_chunks(Rest, [{attributes,Attribs}|Acc]);
parse_chunks([{{"CInf", Size, Chunk} | Rest}, Acc] ->
  <<Bin:Size/binary, _Pad/binary>> = Chunk,
  CInfo = binary_to_term(Bin),
  parse_chunks(Rest, [{compile_info,CInfo}|Acc]);
parse_chunks([{{"LocT", _Size,
  <<_Numberofentries:32/integer, Locals/binary>>} | Rest}, Acc] ->
  <<Locals:_Size/binary>> = Locals,
  parse_chunks(Rest, [{locals,Locals}|Acc]);

```

```

    | Rest], Acc) ->
  parse_chunks(Rest,[{locals,parse_table(Locals)}|Acc]);
parse_chunks([{{"LitT", _ChunkSize,
    <<_CompressedTableSize:32, Compressed/binary>>}
    | Rest], Acc) ->
<<NumLiterals:32,Table/binary>> = zlib:uncompress(Compressed),
Literals = parse_literals(Table),
parse_chunks(Rest,[{literals,Literals}|Acc]);
parse_chunks([{{"Abst", _ChunkSize, <>>} | Rest], Acc) ->
  parse_chunks(Rest,Acc);
parse_chunks([{{"Abst", _ChunkSize, <<AbstractCode/binary>>} | Rest], Acc) ->
  parse_chunks(Rest,[{abstract_code,binary_to_term(AbstractCode)}|Acc]);
parse_chunks([{{"Line", _ChunkSize, <<LineTable/binary>>} | Rest], Acc) ->
<<Ver:32,Bits:32,NumLineInstrs:32,NumLines:32,NumFnames:32,
 Lines:NumLines/binary,Fnames/binary>> = LineTable,
parse_chunks(Rest,[{line,
    [{version,Ver},
     {bits,Bits},
     {num_line_instructions,NumLineInstrs},
     {lines,decode_lineinfo(binary_to_list(Lines),0)},
     {function_names,Fnames}]}|Acc]]);

parse_chunks([Chunk|Rest], Acc) -> %% Not yet implemented chunk
  parse_chunks(Rest, [Chunk|Acc]);
parse_chunks([],Acc) -> Acc.

parse_atoms(<<Atomlength, Atom:Atomlength/binary, Rest/binary>>) when Atomlength > 0->
  [list_to_atom(binary_to_list(Atom)) | parse_atoms(Rest)];
parse_atoms(_Alignment) -> [].

parse_table(<<Function:32/integer,
            Arity:32/integer,
            Label:32/integer,
            Rest/binary>>) ->
  [{Function, Arity, Label} | parse_table(Rest)];
parse_table(<>>) -> [].

parse_code_info(<<Instructionset:32/integer,
                OpcodeMax:32/integer,
                NumberOfLabels:32/integer,
                NumberOfFunctions:32/integer,
                Rest/binary>>) ->
[{instructionset, Instructionset},
 {opcodemax, OpcodeMax},
 {numberoflabels, NumberOfLabels},
 {numberoffunctions, NumberOfFunctions} |
 case Rest of
 <>> -> [];
 _ -> [{newinfo, Rest}]
 end].

```

`parse_literals(<<Size:32,Literal:Size/binary,Tail/binary>>) ->`

```

[binary_to_term(Literal) | parse_literals(Tail)];
parse_literals(<>>) -> [].

-define(tag_i, 1).
-define(tag_a, 2).

decode_tag(?tag_i) -> i;
decode_tag(?tag_a) -> a.

decode_int(Tag,B Bs) when (B band 16#08) =:= 0 ->
    %% N < 16 = 4 bits, NNNN:0:TTT
    N = B bsr 4,
    {{Tag,N},Bs};
decode_int(Tag,B,[]) when (B band 16#10) =:= 0 ->
    %% N < 2048 = 11 bits = 3:8 bits, NNN:01:TTT, NNNNNNNN
    Val0 = B band 2#11100000,
    N = (Val0 bsl 3),
    {{Tag,N},[]};
decode_int(Tag,B Bs) when (B band 16#10) =:= 0 ->
    %% N < 2048 = 11 bits = 3:8 bits, NNN:01:TTT, NNNNNNNN
    [B1|Bs1] = Bs,
    Val0 = B band 2#11100000,
    N = (Val0 bsl 3) bor B1,
    {{Tag,N},Bs1};
decode_int(Tag,B Bs) ->
    {Len,Bs1} = decode_int_length(B,Bs),
    {IntBs,RemBs} = take_bytes(Len,Bs1),
    N = build_arg(IntBs),
    {{Tag,N}},RemBs).

decode_lineinfo([B|Bs], F) ->
    Tag = decode_tag(B band 2#111),
    {{Tag,Num},RemBs} = decode_int(Tag,B,Bs),
    case Tag of
        i ->
            [{F, Num} | decode_lineinfo(RemBs, F)];
        a ->
            [B2|Bs2] = RemBs,
            Tag2 = decode_tag(B2 band 2#111),
            {{Tag2,Num2},RemBs2} = decode_int(Tag2,B2,Bs2),
            [{Num, Num2} | decode_lineinfo(RemBs2, Num2)]
    end;
decode_lineinfo([],_) -> [].

decode_int_length(B, Bs) ->
    {B bsr 5 + 2, Bs}.

take_bytes(N, Bs) ->
    take_bytes(N, Bs, []).

take_bytes(N, [B|Bs], Acc) when N > 0 ->

```

```

take_bytes(N-1, Bs, [B|Acc]);
take_bytes(0, Bs, Acc) ->
{lists:reverse(Acc), Bs}.

```

```

build_arg(Bs) ->
build_arg(Bs, 0).

```

```

build_arg([B|Bs], N) ->
build_arg(Bs, (N bsl 8) bor B);
build_arg([], N) ->
N.

```

```

-module(world).
-export([hello/0]).

```

```
-include("world.hrl").
```

```
hello() -> ?GREETING.
```

```

-module(json_parser).
-export([parse_transform/2]).
```

```
parse_transform(AST, _Options) ->
json(AST, []).
```

```
-define(FUNCTION(Clauses), {function, Label, Name, Arity, Clauses}).
```

```
%% We are only interested in code inside functions.
```

```
json([?FUNCTION(Clauses) | Elements], Res) ->
json(Elements, [?FUNCTION(json_clauses(Clauses)) | Res]);
json([Other|Elements], Res) -> json(Elements, [Other | Res]);
json([], Res) -> lists:reverse(Res).
```

```
%% We are interested in the code in the body of a function.
```

```
json_clauses([{clause, CLine, A1, A2, Code} | Clauses]) ->
[{clause, CLine, A1, A2, json_code(Code)} | json_clauses(Clauses)];
json_clauses([]) -> [].
```

```
-define(JSON(Json), {bin, _, [{bin_element
        , _ , {tuple, _, [Json]}}
        , _ ]})).
```

```
%% We look for: <<"json">> = Json-Term
```

```
json_code([]) -> [];
json_code([?JSON(Json)|MoreCode]) -> [parse_json(Json) | json_code(MoreCode)];
json_code(Code) -> Code.
```

```

%% Json Object -> []
%% Json Array -> List
%% Json String -> <<String>>
%% Json Integer -> Intger
%% Json Float -> Float
%% Variables, should contain Erlang encoded Json
%% parse_json({var, Line, Var})      -> {var, Line, Var};
%% Json Negative Integer or Float
%% parse_json({op, Line, '-', {Type, _, N}}) when Type ==:= integer
%%                                         ; Type ==:= float ->
%%                                         {Type, Line, -N}.
%% parse_json(Code)                  -> io:format("Code: ~p~n",[Code]), Code.

-define(FIELD(Label, Code), {remote, L, {string, _, Label}, Code}).

parse_json_fields([], L) -> {nil, L};
parse_json_fields([?FIELD(Label, Code) | Rest], L) ->
    cons(tuple(str_to_bin(Label, L), parse_json(Code), L),
         , parse_json_fields(Rest, L)
         , L).

tuple(E1, E2, Line)    -> {tuple, Line, [E1, E2]}.
cons(Head, Tail, Line) -> {cons, Line, Head, Tail}.

str_to_bin(String, Line) ->
{bin
, Line
, [{bin_element
, Line
, {string, Line, String}
, default
, default
}
]
}.

```

```
-module(json_test).
-compile({parse_transform, json_parser}).
-export([test/1]).  
  
test(V) ->  
  <<{{  
    "name" : "Jack (\\"Bee\\") Nimble",  
    "format": {  
      "type"      : "rect",  
      "widths"    : [1920,1600],  
      "height"   : (-1080),  
      "interlace" : false,  
      "frame rate": V  
    }  
  }}>>.
```

```
-module(msg).

-export([send_on_heap/0
        ,send_off_heap/0]).  
  
send_on_heap() -> send(on_heap).  
send_off_heap() -> send(off_heap).  
  
send(How) ->  
    %% Spawn a function that loops for a while  
    P2 = spawn(fun () -> receiver(How) end),  
    %% spawn a sending process  
    P1 = spawn(fun () -> sender(P2) end),  
    P1.  
  
sender(P2) ->  
    %% Send a message that ends up on the heap  
    %% {_,S} = erlang:process_info(P2, heap_size),  
    M = loop(0),  
    P2 ! self(),  
    receive ready -> ok end,  
    P2 ! M,  
    %% Print the PCB of P2  
    hipe_bifs:show_pcb(P2),  
    ok.  
  
receiver(How) ->  
    erlang:process_flag(message_queue_data, How),  
    receive P -> P ! ready end,  
    %% loop(100000),  
    receive x -> ok end,  
    P.  
  
loop(0) -> [done];  
loop(N) -> [loop(N-1)].
```

```

-module(stack_machine_compiler).
-export([compile/2]).

compile(Expression, FileName) ->
    [ParseTree] = element(2,
        erl_parse:parse_exprs(
            element(2,
                erl_scan:string(Expression)))),
    file:write_file(FileName, generate_code(ParseTree) ++ [stop()]). 

generate_code({op, _Line, '+', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [add()];
generate_code({op, _Line, '*', Arg1, Arg2}) ->
    generate_code(Arg1) ++ generate_code(Arg2) ++ [multiply()];
generate_code({integer, _Line, I}) -> [push(), integer(I)].

stop()      -> 0.
add()       -> 1.
multiply()  -> 2.
push()      -> 3.
integer(I) ->
    L = binary_to_list(binary:encode_unsigned(I)),
    [length(L) | L].

```

```

#include <stdio.h>
#include <stdlib.h>

char *read_file(char *name) {
    FILE *file;
    char *code;
    long size;

    file = fopen(name, "r");
    if(file == NULL) exit(1);

    fseek(file, 0L, SEEK_END);
    size = ftell(file);
    code = (char*)calloc(size, sizeof(char));
    if(code == NULL) exit(1);

    fseek(file, 0L, SEEK_SET);

    fread(code, sizeof(char), size, file);
    fclose(file);
    return code;
}

#define STOP 0
#define ADD 1
#define MUL 2
#define PUSH 3

```

```

#define pop() (stack[--sp])
#define push(X) (stack[sp++] = X)

int run(char *code) {
    int stack[1000];
    int sp = 0, size = 0, val = 0;
    char *ip = code;

    while (*ip != STOP) {
        switch (*ip++) {
        case ADD: push(pop() + pop()); break;
        case MUL: push(pop() * pop()); break;
        case PUSH:
            size = *ip++;
            val = 0;
            while (size--) { val = val * 256 + *ip++; }
            push(val);
            break;
        }
    }
    return pop();
}

int main(int argc, char *argv[])
{
    char *code;
    int res;

    if (argc > 1) {
        code = read_file(argv[1]);
        res = run(code);
        printf("The value is: %i\n", res);
        return 0;
    } else {
        printf("Give the file name of a byte code program as argument\n");
        return -1;
    }
}

```

```

#include <stdio.h>
#include <stdlib.h>

#define STOP 0
#define ADD 1
#define MUL 2
#define PUSH 3

#define pop() (stack[--sp])
#define push(X) (stack[sp++] = (X))

typedef void (*instructionp_t)(void);

```

```

int stack[1000];
int sp;
instructionp_t *ip;
int running;

void add() { int x,y; x = pop(); y = pop(); push(x + y); }
void mul() { int x,y; x = pop(); y = pop(); push(x * y); }
void pushi(){ int x; x = (int)*ip++; push(x); }
void stop() { running = 0; }

instructionp_t *read_file(char *name) {
    FILE *file;
    instructionp_t *code;
    instructionp_t *cp;
    long size;
    char ch;
    unsigned int val;

    file = fopen(name, "r");
    if(file == NULL) exit(1);

    fseek(file, 0L, SEEK_END);
    size = ftell(file);
    code = calloc(size, sizeof(instructionp_t));
    if(code == NULL) exit(1);
    cp = code;

    fseek(file, 0L, SEEK_SET);
    while ( ( ch = fgetc(file) ) != EOF )
    {
        switch (ch) {
        case ADD: *cp++ = &add; break;
        case MUL: *cp++ = &mul; break;
        case PUSH:
            *cp++ = &pushi;
            ch = fgetc(file);
            val = 0;
            while (ch--) { val = val * 256 + fgetc(file); }
            *cp++ = (instructionp_t) val;
            break;
        }
    }
    *cp = &stop;

    fclose(file);
    return code;
}

int run() {
    sp = 0;
    running = 1;
}

```

```

while (running) (*ip++)();

return pop();
}

int main(int argc, char *argv[])
{
    if (argc > 1) {
        ip = read_file(argv[1]);
        printf("The value is: %i\n", run());
        return 0;
    } else {
        printf("Give the file name of a byte code program as argument\n");
        return -1;
    }
}

```

```

-module(share).

-export([share/2, size/0]).

share(0, Y) -> {Y,Y};
share(N, Y) -> [share(N-1, [N|Y]) || _ <- Y].

size() ->
    T = share:share(5,[a,b,c]),
    {{size, erts_debug:size(T)},
     {flat_size, erts_debug:flat_size(T)}}.

```

```

-module(send).

-export([test/0]).

test() ->
    P2 = spawn(fun() -> p2() end),
    P1 = spawn(fun() -> p1(P2) end),
    {P1, P2}.

p2() ->
    receive
        M -> io:format("P2 got ~p", [M])
    end.

p1(P2) ->
    L = "hello",
    M = {L, L},
    P2 ! M,
    io:format("P1 sent ~p", [M]).

```

```

-module(lb).
-export([start/0]).

start() ->
    Workers = [spawn(fun worker/0) || _ <- lists:seq(1,10)],
    LoadBalancer = spawn(fun() -> loop(Workers, 0) end),
    {ok, Files} = file:list_dir("."),
    Loaders = [spawn(fun() -> loader(LoadBalancer, F) end) || F <- Files],
    {Loaders, LoadBalancer, Workers}.

loader(LB, File) ->
    case file:read_file(File) of
        {ok, Bin} -> LB ! Bin;
        _Dir -> ok
    end,
    ok.

worker() ->
    receive
        Bin ->
            io:format("Byte Size: ~w~n", [byte_size(Bin)]),
            garbage_collect(),
            worker()
    end.

loop(Workers, N) ->
    receive
        WorkItem ->
            Worker = lists:nth(N+1, Workers),
            Worker ! WorkItem,
            loop(Workers, (N+1) rem length(Workers))
    end.

```

show.

```

-module(show).
-export([ hex_tag/1
        , tag/1
        , tag_to_type/1
        ]).

tag(Term) ->
    Bits = integer_to_list(erlang:system_info(wordsize)*8),
    FormatString = "~" ++ Bits ++ ".2.0B",
    io:format(FormatString,[hipe_bifs:term_to_word(Term)]).

hex_tag(Term) ->
    Chars = integer_to_list(erlang:system_info(wordsize)*2),
    FormatString = "~" ++ Chars ++ ".16.0b",
    io:format(FormatString,[hipe_bifs:term_to_word(Term)]).

tag_to_type(Word) ->
    case Word band 2#11 of
        2#00 -> header;
        2#01 -> cons;
        2#10 -> boxed;
        2#11 ->
            case (Word bsr 2) band 2#11 of
                2#00 -> pid;
                2#01 -> port;
                2#10 ->
                    case (Word bsr 4) band 2#11 of
                        00 -> atom;
                        01 -> 'catch';
                        10 -> 'UNUSED';
                        11 -> nil
                    end;
                2#11 -> smallint
            end
    end
end.

```

```

diff --git a/erts/emulator/hipe/hipe_debug.c b/erts/emulator/hipe/hipe_debug.c
index ace4894..7a888cc 100644
--- a/erts/emulator/hipe/hipe_debug.c
+++ b/erts/emulator/hipe/hipe_debug.c
@@ -39,16 +39,16 @@
 #include "hipe_debug.h"
 #include "erl_map.h"

-static const char dashes[2*sizeof(long)+5] = {
-    [0 ... 2*sizeof(long)+3] = '_'
+static const char dashes[2*sizeof(long *)+5] = {
+    [0 ... 2*sizeof(long *)+3] = '_'
};


```

```

-static const char dots[2*sizeof(long)+5] = {
-    [0 ... 2*sizeof(long)+3] = '.'
+static const char dots[2*sizeof(long *)+5] = {
+    [0 ... 2*sizeof(long *)+3] = '.'
};

-static const char stars[2*sizeof(long)+5] = {
-    [0 ... 2*sizeof(long)+3] = '*'
+static const char stars[2*sizeof(long *)+5] = {
+    [0 ... 2*sizeof(long *)+3] = '*'
};

extern Uint beam_apply[];
@@ -56,52 +56,56 @@ extern Uint beam_apply[];
static void print_beam_pc(BeamInstr *pc)
{
    if (pc == hipe_beam_pc_return) {
-    printf("return-to-native");
+    erts_printf("return-to-native");
    } else if (pc == hipe_beam_pc_throw) {
-    printf("throw-to-native");
+    erts_printf("throw-to-native");
    } else if (pc == &beam_apply[1]) {
-    printf("normal-process-exit");
+    erts_printf("normal-process-exit");
    } else {
        BeamInstr *mfa = find_function_from_pc(pc);
        if (mfa)
            erts_printf("%T:%T/%bpu + 0x%bp",
-                       mfa[0], mfa[1], mfa[2], pc - &mfa[3]);
+                       mfa[0], mfa[1], mfa[2], pc - &mfa[3]);
        else
-            printf("?");
+            erts_printf("?");
    }
}

static void catch_slot(Eterm *pos, Eterm val)
{
    BeamInstr *pc = catch_pc(val);
-    printf(" | 0x%0*lx | 0x%0*lx | CATCH 0x%0*lx (BEAM ",
+    erts_printf(" | 0x%0*lx | 0x%0*lx | CATCH 0x%0*lx",
        2*(int)sizeof(long), (unsigned long)pos,
        2*(int)sizeof(long), (unsigned long)val,
        2*(int)sizeof(long), (unsigned long)pc);
+    erts_printf("\r\n");
+    erts_printf(" | %*s | %*s | (BEAM ",
+                2*(int)sizeof(long), " ",
+                2*(int)sizeof(long), " ");
        print_beam_pc(pc);
-    printf("\r\n");
+    erts_printf("\r\n");
}
}

static void print_beam_cp(Eterm *pos, Eterm val)

```

```

{
- printf(" |%s|%s| BEAM ACTIVATION RECORD\r\n", dashes, dashes);
- printf(" | 0x%0*lx | 0x%0*lx | BEAM PC ",
+ erts_printf(" |%s|%s| BEAM ACTIVATION RECORD\r\n", dashes, dashes);
+ erts_printf(" | 0x%0*lx | 0x%0*lx | BEAM PC ",
    2*(int)sizeof(long), (unsigned long)pos,
    2*(int)sizeof(long), (unsigned long)val);
print_beam_pc(cp_val(val));
- printf("\r\n");
+ erts_printf("\r\n");
}

static void print_catch(Eterm *pos, Eterm val)
{
- printf(" |%s|%s| BEAM CATCH FRAME\r\n", dots, dots);
+ erts_printf(" |%s|%s| BEAM CATCH FRAME\r\n", dots, dots);
catch_slot(pos, val);
- printf(" |%s|%s|\r\n", stars, stars);
+ erts_printf(" |%s|%s|\r\n", stars, stars);
}

static void print_stack(Eterm *sp, Eterm *end)
{
- printf(" | %*s | %*s |\r\n",
+ erts_printf(" | %*s | %*s |\r\n",
    2+2*(int)sizeof(long), "Address",
    2+2*(int)sizeof(long), "Contents");
while (sp < end) {
@@ -111,56 +115,68 @@ static void print_stack(Eterm *sp, Eterm *end)
    else if (is_catch(val))
        print_catch(sp, val);
    else {
-        printf(" | 0x%0*lx | 0x%0*lx | ",
+        erts_printf(" | 0x%0*lx | 0x%0*lx | ",
            2*(int)sizeof(long), (unsigned long)sp,
            2*(int)sizeof(long), (unsigned long)val);
        erts_printf("%.30T", val);
-        printf("\r\n");
+        erts_printf("\r\n");
    }
    sp += 1;
}
- printf(" |%s|%s|\r\n", dashes, dashes);
+ erts_printf(" |%s|%s|\r\n", dashes, dashes);
}

void hipe_print_estack(Process *p)
{
- printf(" |      BEAM STACK      |\r\n");
+ erts_printf(" |      BEAM STACK      |\r\n");
    print_stack(p->stop, STACK_START(p));
}

static void print_heap(Eterm *pos, Eterm *end)

```

```

{
- printf("From: 0x%0*lx to 0x%0*lx\n\r",
-     2*(int)sizeof(long), (unsigned long)pos,
-     2*(int)sizeof(long), (unsigned long)end);
- printf(" | H E A P | \r\n");
- printf(" | %*s | %*s | \r\n",
-     2+2*(int)sizeof(long), "Address",
-     2+2*(int)sizeof(long), "Contents");
- printf(" | %s| %s| \r\n", dashes, dashes);
+ erts_printf("From: 0x%0*lx to 0x%0*lx\n\r",
+     2*(int)sizeof(long *), (unsigned long)pos,
+     2*(int)sizeof(long *), (unsigned long)end);
+ erts_printf(" | %*s%*s%*s | \r\n",
+     2+1*(int)sizeof(long), " ",
+     2+1*(int)sizeof(long), "H E ",
+     3, "A P",
+     2*(int)sizeof(long), " "
+ );
+ erts_printf(" | %*s | %*s | \r\n",
+     2+2*(int)sizeof(long *), "Address",
+     2+2*(int)sizeof(long *), "Contents");
+ erts_printf(" | %s| %s| \r\n", dashes, dashes);
    while (pos < end) {
        Eterm val = pos[0];
-        printf(" | 0x%0*lx | 0x%0*lx | ",
-            2*(int)sizeof(long), (unsigned long)pos,
-            2*(int)sizeof(long), (unsigned long)val);
+        if ((is_arity_value(val)) || (is_thing(val))) {
+            erts_printf(" | 0x%0*lx | 0x%0*lx | ",
+                2*(int)sizeof(long *), (unsigned long)pos,
+                2*(int)sizeof(long *), (unsigned long)val);
+        } else {
+            erts_printf(" | 0x%0*lx | 0x%0*lx | ",
+                2*(int)sizeof(long *), (unsigned long)pos,
+                2*(int)sizeof(long *), (unsigned long)val);
+            erts_printf("%-*.*T", 2*(int)sizeof(long),(int)sizeof(long), val);
+
+        }
+        ++pos;
+        if (is_arity_value(val))
-            printf("Arity(%lu)", arityval(val));
+            erts_printf("Arity(%lu)", arityval(val));
        else if (is_thing(val)) {
            unsigned int ari = thing_arityval(val);
-            printf("Thing Arity(%u) Tag(%lu)", ari, thing_subtag(val));
+            erts_printf("Thing Arity(%u) Tag(%lu)", ari, thing_subtag(val));
            while (ari) {
-                printf("\r\n | 0x%0*lx | 0x%0*lx | THING",
-                    2*(int)sizeof(long), (unsigned long)pos,
-                    2*(int)sizeof(long), (unsigned long)*pos);
+                erts_printf("\r\n | 0x%0*lx | 0x%0*lx | THING",
+                    2*(int)sizeof(long *), (unsigned long)pos,
+                    2*(int)sizeof(long *), (unsigned long)*pos);
                ++pos;
            }
        }
    }
}

```

```

--ari;
}

- } else
-     erts_printf("%.30T", val);
- printf("\r\n");
+ }
+ erts_printf("\r\n");
}
- printf(" |$|$|\r\n", dashes, dashes);
+ erts_printf(" |$|$|\r\n", dashes, dashes);
}

void hipe_print_heap(Process *p)
@@ -170,74 +186,85 @@ void hipe_print_heap(Process *p)

void hipe_print_pcb(Process *p)
{
-     printf("P: 0x%0*lx\r\n", 2*(int)sizeof(long), (unsigned long)p);
-     printf("-----\r\n");
-     printf("Offset| Name      | Value      | *Value      |\r\n");
+     erts_printf("P: 0x%0*lx\r\n", 2*(int)sizeof(long *), (unsigned long)p);
+     erts_printf("-----%s\r\n", dashes, dashes);
+     erts_printf("Offset| Name      | %*s | %*s |\r\n",
+                 2*(int)sizeof(long *), "Value",
+                 2*(int)sizeof(long *), "*Value"
+ );
#undef U
#define U(n,x) \
-     printf(" % 4d | %s | 0x%0*lx |           |\r\n", (int)offsetof(Process,x), n,
2*(int)sizeof(long), (unsigned long)p->x)
+     erts_printf(" % 4d | %s | 0x%0*lx | %*s | \r\n", (int)offsetof(Process,x), n, 2*(int)sizeof
(long *), (unsigned long)p->x, 2*(int)sizeof(long *), " ")
#undef P
#define P(n,x) \
-     printf(" % 4d | %s | 0x%0*lx | 0x%0*lx |\r\n", (int)offsetof(Process,x), n, 2*(int)sizeof(long),
(unsigned long)p->x, 2*(int)sizeof(long), p->x ? (unsigned long)*(p->x) : -1UL)
+     erts_printf(" % 4d | %s | 0x%0*lx | 0x%0*lx |\r\n", (int)offsetof(Process,x), n, 2*(int)sizeof
(long *), (unsigned long)p->x, 2*(int)sizeof(long *), p->x ? (unsigned long)*(p->x) : -1UL)

-     U("htop      ", htop);
-     U("hend      ", hend);
-     U("heap      ", heap);
-     U("heap_sz   ", heap_sz);
-     U("stop      ", stop);
-     U("gen_gcs   ", gen_gcs);
-     U("max_gen_gcs", max_gen_gcs);
-     U("high_water ", high_water);
-     U("old_hend   ", old_hend);
-     U("old_htop   ", old_htop);
-     U("old_head   ", old_heap);
-     U("min_heap_..", min_heap_size);
-     U("rcount     ", rcount);
-     U("id         ", common.id);
-     U("reds       ", reds);

```

```

- U("tracer      ", common.tracer);
- U("trace_flag.", common.trace_flags);
- U("group_lead.", group_leader);
- U("flags       ", flags);
- U("fvalue      ", fvalue);
- U("freason     ", freason);
- U("fcalls      ", fcalls);
+ U("id          ", common.id);
+ U("htop        ", htop);
+ U("hend        ", hend);
+ U("heap         ", heap);
+ U("heap_sz     ", heap_sz);
+ U("stop         ", stop);
+ U("gen_gcs     ", gen_gcs);
+ U("max_gen_gcs", max_gen_gcs);
+ U("high_water  ", high_water);
+ U("old_hend    ", old_hend);
+ U("old_htop    ", old_htop);
+ U("old_head    ", old_heap);
+ U("min_heap_size", min_heap_size);
+ U("msg.first   ", msg.first);
+ U("msg.last    ", msg.last);
+ U("msg.save    ", msg.save);
+ U("msg.len     ", msg.len);
#ifndef ERTS_SMP
+ U("msg_inq.first", msg_inq.first);
+ U("msg_inq.last ", msg_inq.last);
+ U("msg_inq.len  ", msg_inq.len);
#endif
+ U("mbuf        ", mbuf);
+ U("mbuf_sz     ", mbuf_sz);
+ U("rcount       ", rcount);
+ U("reds         ", reds);
+ U("tracer      ", common.tracer);
+ U("trace_flags ", common.trace_flags);
+ U("group_leader", group_leader);
+ U("flags       ", flags);
+ U("fvalue      ", fvalue);
+ U("freason     ", freason);
+ U("fcalls      ", fcalls);
/*XXX: ErlTimer tm; */
- U("next        ", next);
+ U("next        ", next);
/*XXX: ErlOffHeap off_heap; */
- U("reg         ", common.u.alive.reg);
- U("nlinks      ", common.u.alive.links);
/*XXX: ErlMessageQueue msg; */
- U("mbuf        ", mbuf);
- U("mbuf_sz     ", mbuf_sz);
- U("dictionary  ", dictionary);
- U("seq..clock  ", seq_trace_clock);
- U("seq..astcnt", seq_trace_lastcnt);
- U("seq..token  ", seq_trace_token);
- U("initial[0]  ", u.initial[0]);

```

```

- U("intial[1] ", u.initial[1]);
- U("intial[2] ", u.initial[2]);
- P("current ", current);
- P("cp ", cp);
- P("i ", i);
- U("catches ", catches);
- U("arity ", arity);
- P("arg_reg ", arg_reg);
- U("max_arg_reg", max_arg_reg);
- U("def..reg[0]", def_arg_reg[0]);
- U("def..reg[1]", def_arg_reg[1]);
- U("def..reg[2]", def_arg_reg[2]);
- U("def..reg[3]", def_arg_reg[3]);
- U("def..reg[4]", def_arg_reg[4]);
- U("def..reg[5]", def_arg_reg[5]);
+ U("reg ", common.u.alive.reg);
+ U("nlinks ", common.u.alive.links);
+ U("dictionary ", dictionary);
+ U("seq...clock ", seq_trace_clock);
+ U("seq...astcnt ", seq_trace_lastcnt);
+ U("seq...token ", seq_trace_token);
+ U("intial[0] ", u.initial[0]);
+ U("intial[1] ", u.initial[1]);
+ U("intial[2] ", u.initial[2]);
+ P("current ", current);
+ P("cp ", cp);
+ P("i ", i);
+ U("catches ", catches);
+ U("arity ", arity);
+ P("arg_reg ", arg_reg);
+ U("max_arg_reg", max_arg_reg);
+ U("def..reg[0]", def_arg_reg[0]);
+ U("def..reg[1]", def_arg_reg[1]);
+ U("def..reg[2]", def_arg_reg[2]);
+ U("def..reg[3]", def_arg_reg[3]);
+ U("def..reg[4]", def_arg_reg[4]);
+ U("def..reg[5]", def_arg_reg[5]);
#ifdef HIPE
- U("nsp ", hipe.nsp);
- U("nstack ", hipe.nstack);
- U("nstend ", hipe.nstend);
- U("ncallee ", hipe.u.ncallee);
+ U("nsp ", hipe.nsp);
+ U("nstack ", hipe.nstack);
+ U("nstend ", hipe.nstend);
+ U("ncallee ", hipe.u.ncallee);
    hipe_arch_print_pcb(&p->hipe);
#endif /* HIPE */
#undef U
#undef P
- printf("-----\r\n");
+ erts_printf("-----%s%s\r\n", dashes, dashes);
}

```