The Safe and Secure Virtual Machine - Sense-VM architecture

Author: Bo Joel Svensson, Abhiroop Sarkar

WORK IN PROGRESS - DRAFT

Todo: add something about mailboxes, message passing

Sense-VM is a virtual machine for IoT and embedded applications in general. Sense-VM is a bytecode virtual machine (think Java-VM, not an operating system hypervisor) that provides a base level of safety, robustness and security.

Sense-VM targets 32/64Bit microcontroller based systems with at least about 128Kb of Read Write Memory (RWM).

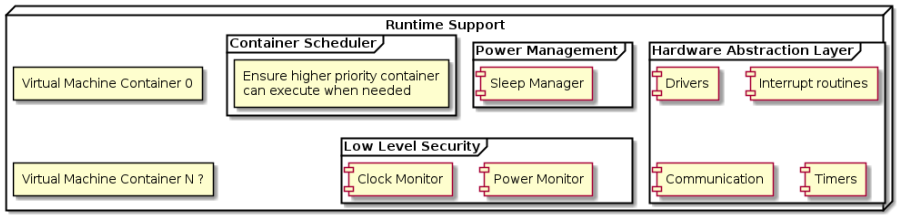
Sense-VM consists of a runtime-system for execution of compiled linearly-stored bytecode within isolated containers. The bytecode programs cannot mutate arbitrary memory addresses and all accesses to underlying hardware goes via the runtime system.

Sense-VM can concurrently run a number of isolated virtual exection environments, called Virtual Machine Containers. Each virtual machine container contains the resources needed to execute one program and consist of registers and memory that is private to that environment. A program running within a virtual machine container could potentially be concurrent itself, running a cooperative scheduler. A criticality level (priorities) can be associated with each virtual machine container. Peripherals and external hardware resources are assigned to a single virtual machine container to rule out competition for resources between applications of different criticality (priority).

Sense-VM is implemented in standard C for portability and static analysers, infer (from Facebook) and scan-build (from the Clang framework) are used to detect problematic code.

Sense-VM specifies an interface downwards towards hardware functionality and peripherals, but implemnetation of that interface from below is not considered a part of Sense-VM per se. Microcontroller platforms come in very different forms so it makes sense to leave the low-level implementation to the specific use case. We do however provide implementations based on ChibiOS (Or contikiOs or ST HAL, or so on..) *Todo: Such interfaces downwards towards the hardware are not yet specified) Todo: No such "ref" implementation is implemented or specified)*

The Sense-VM Virtual Machine



Hardware Abstraction Layer

The hardware abstraction layer (HAL) will probably be implemented by building upon an existing HAL system such as ChibiOS, ContikiOs or ZephyrOS. Other HALs could also be candidates: CMSIS, HAL from ST for STM32 platforms for example. We should try to keep as much of the code as possible portable so that it can be ported to HALs that parhaps have more desirable features. The HAL will be a collection of functions (and perhaps routines running on a timer interrupt periodically if needed) that can be called from the runtime support system.

Sleep Manager

If there is no processing going on (reported from the various schedulers) the sleep manager puts the system to sleep for a period of time such that it wakes up when it is time for the context/container with the earliest "wake-up" time to start executing.

The sleep manager may also be prompted to wake the system up by for example a sensor driver when the sensor has new samples to process.

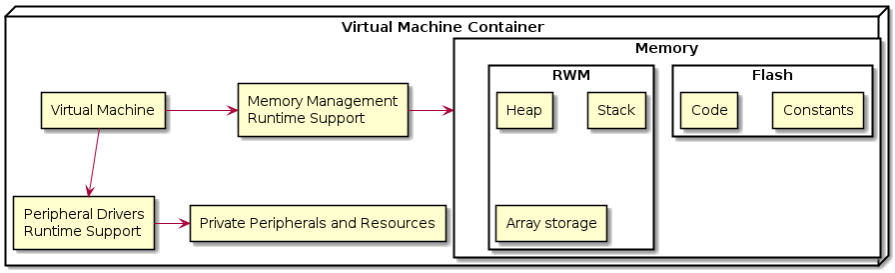
It is not unthinkable that there are applications that will just sleep indefinitely unless there is outside stimulus. It may still be desirable that these applications wake up periodically and just let the monitoring system know that it is still alive and ok (a heartbeat).

Container Scheduler and Virtual Machine Containers

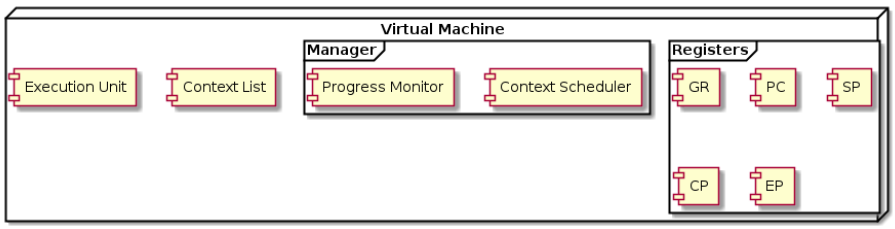
A virtual machine container is an isolated executing environment for a virtual machine.

The container scheduler ensures that each container gets a time slot to execute. Possibly, these containers could be implemented using a "thread" feature of an underlying OS/RTOS/HAL or be given a certain number of iterations of some main loop.

Containers should be isolated from eachother to allow different level of criticality to different containers. High criticality containers should be prioritised over low criticality containers.



A VM container consist of a virtual machine instance and a set of dedicated and private resources.



Virtual Machine

The virtual machine is based upon the Categorical Abstract Machine (CAM) and consists of number of registers and and an execution unit. (More details later)

Registers

- GR - General register. Used ot hold intermediate values, results and one of the arguments in a function call.
- PC - Program Counter. Instruction to execute. Index into code memory.
- SP - Pointer to stack data structure current in use (not a traditional stack pointer)
- CP - Pointer to location in code area. The base address for the PC indices.
- EP - Pointer to an environment datastructure.

Execution Unit

The execution unit reads instructions from code memory at location stored in PC register and evaluates each instruction over the state.

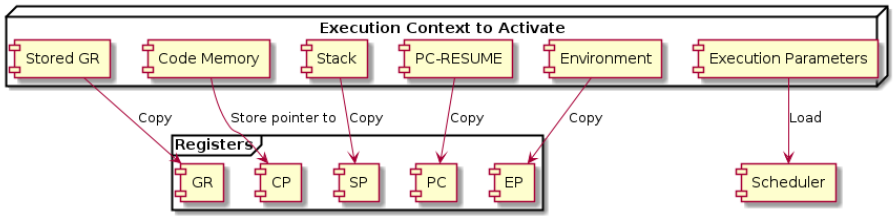
Context Scheduler and Context List

There is a list of contexts (separate instances of code and state) that can execute in a time-shared fashion on the execution unit.

Todo: Maybe cooperative scheduling at this level? This would require "go to sleep" operations in the bytecode.

The contexts that are executed on one VM are sharing the memory resources of the Virtual Machine Container.

Context Switching



The context list consists of some number of "Execution Context to Activate". A context is activated by copying the stored register state into the VM. Putting a context to sleep works in the reversed way.

Memory subsystem

Each Virtual Machine Container (VMC) has access to a private memory subsystem consisting of a garbage collected heap, array storage memory, constants memory and code memory.

Todo: The current garbage collected heap is a placeholder for now. Later we plan ot experiment with different kinds of managed memory subsystems that may be better suited for embedded and IoT.

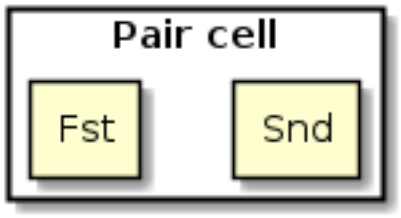
Garbage collected heap

The garbage collected heap is a supply of pairs. The pairs can contain values, pointers to arrays in the Array store, pointers to values in the constants memory or pointers into the heap (for the purpose of building linked structures).

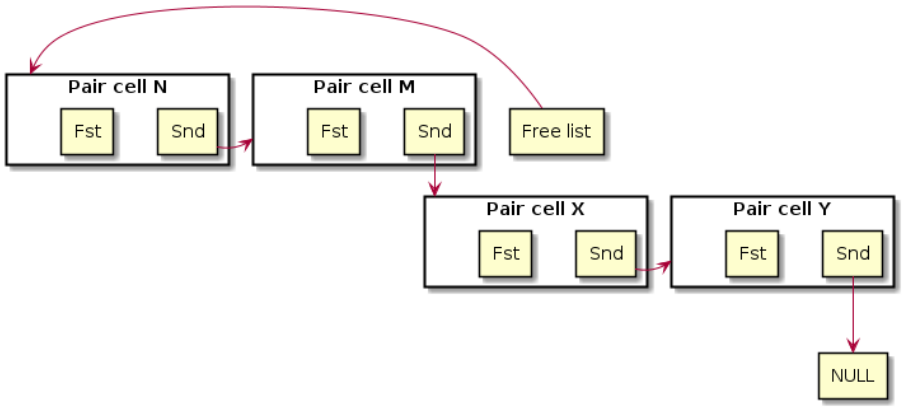
The heap consists of a fixed numer of pairs that is decided at compilation time for each VMC.

Unused pairs are linked together into a structure referred to as the *free list*.

Each pair cell consists of 2 bit quantities:



The free list is linked up at the *snd* component.



The value -1 (negative one) or 4294967295 depending on how you want to view it, is the representative of NULL when it comes to heap pointers.

Array storage

Constants memory

Code memory

Scheduling

Virtual Machine Container Scheduling

Priority based scheduling with pre-emption between virtual machine containers. The VMC scheduling is implemented on top of threading functionality of the underlying HAL. If there are no threading abstractions in the HAL, the scheduling can be implemented using a periodic timer interrupt or explicitly allocating a number of iterations of a main-loop to each VMC.

Cooperative scheduling within a Virtual Machine

A program may consist of a number of cooperating tasks that are never pre-empted. The running task gives up the Execution Unit by executing `sleep ns` instruction.

The Categorical Abstract Machine

Instruction set

To save space in the instr-set, constants are stored in a constants pool (Constants memory). Instructions such as `LOADI` (that takes stores an integer value into the value register) takes and index into an array of integers stored in the constants memory.

OpCode Mnemonic	Value	Arguments	Size (Bytes - including arrguments)
FST	0x00	0	1
SND	0x01	0	1
ACC <n>	0x02	1	2
REST <n>	0x03	1	2
PUSH	0x04	0	1
SWAP	0x05	0	1
LOADI <i>	0x06	1	3
LOADB <b>	0x07	1	2
CLEAR	0x08	0	1
CONS	0x09	0	1
CUR <l>	0x0A	1	3
PACK <t>	0x0B	1	3
SKIP	0x0C	0	1
STOP	0x0D	0	1
APP	0x0E	0	1
RETURN	0x0F	0	1
CALL <l>	0x10	1	3
GOTO <l>	ox11	1	3
GOTOFALSE <l>	0x12	1	3
SWITCH <n> $<t_1>$ $<l_1>$ .. $<t_n>$ $<l_n>$	0x13	1 + 2n	2 + 4n (max n = 256)
ABS	0x14	0	1
NEG	0x15	0	1
NOT	0x16	0	1
DEC	0x17	0	1
ADDI	0x18	0	1
MULI	0x19	0	1
MINI	0x1A	0	1
ADDF	0x1B	0	1
MULF	0x1C	0	1
MINF	0x1D	0	1
GT	0x1E	0	1
LT	0x1F	0	1
EQ	0x20	0	1
GE	0x21	0	1

OpCode Mnemonic	Value	Arguments	Size (Bytes - including arrguments)
LE	0x22	0	1

- <n> - Positive ints - 1 byte long
- <l> - Positive ints for label numbers - 2 bytes long
- <b> - Boolean 1 byte long; 7 bits wasted
- <t> - Tag for a constructor - 2 bytes long
- <i> - index from int pool - max_index_size = 65536. The int itself can be upto 4 bytes long

FIXME: Currently we use the string pool for tags and not some fixed size 2 bytes tag

Bytecode format

TODO: We should also make room in the bytecode format for checksums

Bytecode file contents	Explanation
FE ED CA FE	*Magic Number* - 4 bytes
FF	*Version of bytecode* - 1 byte;
00 03	*Int Pool count* - 2 bytes - 65536 ints possible
00 0E ED 24	Example integer `978212`, size upto 4 bytes
00 00 CE 35	Example integer `52789`
00 01 C4 4D	Example integer `115789`
...	
00 0A	*String Pool count* - Each byte is indexed unlike Int Pool where every 4 bytes is an index.
48 65 6c 6c 6f	Example string “Hello”
57 6f 72 6c 64	Example string “World”
..	
00 00	*Native Pool count* - 2 bytes - index for native functions
..	
00 00 00 FF	Code Length - Max size of 4 bytes Code
.	
.	

Security Measures

TODO: This needs to be thought out parhaps after doing some kind of a Threat Analysis

Fault Tolerance and Reliability

TODO: Try to build in a base set of functionality for fault tolerance and reliability

Communication

Between contexts within a Virtual Machine Container

Contexts executing within a virtual machine container share a heap. Data can easily be exchanged between tasks if both tasks know of a common name referencing a heap structure. Since the Heap is private to the container and contexts never pre-empt eachother, no locking mechanism should be needed for accesses to these shared structures.

Between Virtual Machine Containers

Low criticality containers should not be able to influence the execution of a higher criticality container in any way.
Todo: What mechanisms for communication do we need to add for Container -> Container communication?

Thoughts

Splitting concurrency up between internal to VM container via contexts and external between containers open up to running VM containers on different cores in parallel.

The Concurrency between different containers may allow mixed criticality applications sense each application will execute in an isolated memory and with access only to private resources.

Management of other resources is also important. Communication interfaces, sensors etc. I think it would be beneficial to assign such resources to a VM Container and never allow the same interface to be connected to more than one VM Container.

This is needed in the HAL layer for safety.

- How is volatile memory handled secured : Double storage or memory check
  - Everything stored twice once bitreversed.
  - The smaller "doubling" the worse.
  - Checkerboard memory testing.
- Check on flash memory constants etc, CRC checksum (CRC Signature of flash memory) CCIT
- Sequence-Control: a token that is incremented.
  - Connected to a watch-dog. Check that token only takes valid values.
  - What about Erlang like supervisor processes?
    - Look at Erlang handling of failures.
- Redundancy mechinism 3 way voting etc?

Concurrency

Here we are talking about concurrency within a single VM container. Each container is made up of a statically fixed number of contexts. Each context has its own stack and an environment register. There is common heap owned by the container. When starting a new context, there are 2 options for allocating the stack memory -

- It might be allocated from the heap owned by the container.
- If it is possible to statically estimate the number of threads then that much stack memory(for each thread) can be statically allocated on a separate area as well.

The parent context can use the stack memory of the container to initialize the stack.

We can impose some kind of channel abstraction for communication between the threads.

Hardware abstraction and Drivers (Bridge)

An API for driver implementation that has a set of high-level interfacing functions and a set of low-level interfacing functions.

This interface has to be able to handle many different kinds of peripherals that operate in very different ways.

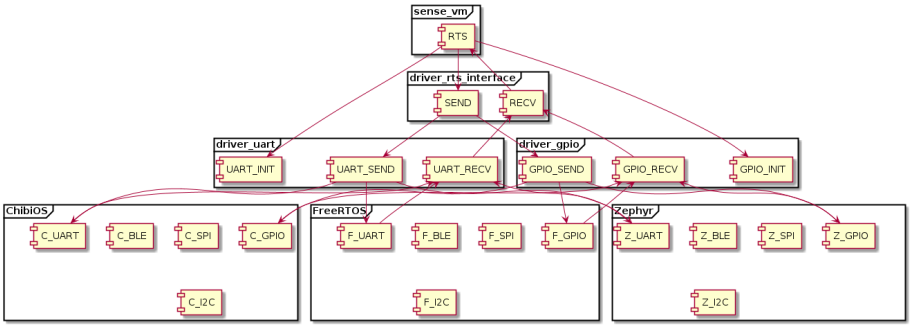
High-level API thought.

```
/* RTS interface - interfaces with the scheduler */
typedef struct {

    volatile bool rdy_recv; /* driver is ready to receive (buffer is not full) */
    volatile bool rdy_send; /* driver is ready to send (buffer is not empty) */

    cam_value_t (*recv)();  /* CAM values or something else? */
    bool (*send)(cam_value_t); /* CAM values or something else? */

} driver_rts_if_t;

/* Each driver provides a RTS Function for initialization */
bool init_X_driver(driver_rts_if_t, more parameters);

/* Example */
bool init_uart_driver(driver_rts_if_t *drv, more parameters );


/* Possible alternative for DMA drivers */
bool init_driver_dma(driver_rts_if_t *drv, uint8_t *array); /* may need additional parameter
```

Init driver takes a pointer to a `driver_rts_if_t` rather than returns one. This is so that the RTS is the entity in control of where and how to store these `driver_rts_if_t` data structures. One thought is that the RTS will have an array of a fixed number of `driver_rts_if_t` storage locations.

In the case of a DMA driver, the RTS needs to know how to sensibly turn the raw array into CAM values that make sense.

I do not think that the DMA engine should be writing data directly to "array" it should have a driver private buffer that it writes to. Each time the DMA "half-full" or "full" interrupt occurs, data should be copied from the private buffer to the "array".



The picture above is simplifying the problem a bit. There is also an additional layer of architectures and "boards". It would be good to be able to leverage on some existing effort here and the best effort currently is probably ZephyrOS.

OS	PROS	CONS
Zephyr	Supports lots of boards	Huge
ChibiOS	Simple and familiar	Very STM focused
FreeRTOS	Well established	No HAL

N times M problem