Operating System Monitor Application (OS_MON)

version 1.3

Joe Armstrong

1997-05-02

Typeset in LaTeX from SGML source using the DOCBUILDER 3.0 Document System.

Contents

OS_MON Reference Manual

Short Summaries

- Application os_mon [page 4] – OS Monitoring Application
- Erlang Module cpu_sup [page 6] – A CPU Load Supervisor Process
- Erlang Module disksup [page 8] – A Disk Supervisor Process.
- Erlang Module memsup [page 10] – A memory Supervisor Process
- Erlang Module nteventlog [page 13] – This module implements a generic interface to the WIndows NT event log. The module is specific to Windows NT and in some ways replace the os_sup module, which is highly Unix specific.
- Erlang Module os_sup [page 16] – This module, together with some dedicated UNIX processes, implements a message passing service from the Solaris operating system to the error logger in the Erlang system. The Solaris (SunOS 5.x) messages are retrieved from the syslog-daemon, syslogd.

os_mon

No functions are exported

cpu_sup

The following functions are exported:

- `nprocs() -> UnixProcesses`
  [page 7] Gets the number of UNIX processes running on this host
- `avg1() -> SystemLoad`
  [page 7] Gets the system load average from the last minute
- `avg5() -> SystemLoad`
  [page 7] Gets the system load average for the last five minutes
- `avg15() -> SystemLoad`
  [page 7] Gets the system load average for the last fifteen minutes

disksup

The following functions are exported:

- `get_check_interval() -> Time`
  [page 8] How often,in milliseconds, the disks are checked
- `get_disk_data() -> [DiskData]`
  [page 8] Gets data for the disks in the system
- `get_almost_full_threshold() -> integer()`
  [page 8]

memsup

The following functions are exported:

- `get_check_interval() -> Time`
  [page 11] How often (in milliseconds) memory is checked
- `get_memory_data() -> MemData`
  [page 11] Gets data for the memory in the system
- `get_system_memory_data() -> MemDataList`
  [page 11] Gets system dependent memory data
- `get_procmem_high_watermark() -> integer()`
  [page 12]
- `get_sysmem_high_watermark() -> integer()`
  [page 12]

nteventlog

The following functions are exported:

- `start(Identifier,MFA) -> Result`
  [page 13] Starts the NT eventlog server
- `start_link(Identifier,MFA) -> Result`
  [page 14] Starts and links the NT eventlog server
- `stop() -> Result`
  [page 14] Stops the message passing service

os_sup

The following functions are exported:

- `start() -> Result`
  [page 16] Starts the message passing service
- `start_link() -> Result`
  [page 16] Starts and links the message passing service
- `stop() -> Result`
  [page 17] Stops the message passing service

os_mon (Application)

This section describes the os_mon application in Erlang. The OS Monitoring application provides the following services:

- cpu_sup
- disksup
- memsup
- os_sup

Configuration

The following configuration parameters are defined for the OS Monitoring application. Refer to application(3) for more information about configuration parameters.

start_disksup = bool() <optional> Specifies if disksup should be started. The default is true.

start_memsup = bool() <optional> Specifies if memsup should be started. The default is true.

start_os_sup = bool() <optional> Specifies if os_sup should be started. The default is false.

disk_space_check_interval = integer() <optional> Defines how often, in minutes, the disksup process should check the disk space. The default is 30 minutes.

disk_almost_full_threshold = float() <optional> Defines what percentage of total disk space can be utilized before the disk_almost_full alarm is set. The default is 0.80 (80%).

memory_check_interval = integer() <optional> Defines how often, in minutes, the memsup process should check the memory. The default is one minute.

system_memory_high_watermark = float() <optional> Defines what percentage of the available system memory can be allocated before the corresponding alarm is set. The default is 0.8 (80%).

process_memory_high_watermark = float() <optional> Defines what percentage of the available system memory can be allocated by one Erlang process before the corresponding alarm is set. The default is 0.05 (5%).

os_sup_own = string() Defines which directory contains the backup copy and the Erlang specific configuration files for syslogd, and the named pipe to receive the messages from syslogd.
Usually, this parameter has the value “/etc”.

`os_sup_syslogconf = string()` Defines the full file name of the configuration file for `syslogd`.
Usually, this parameter has the value "/etc/syslog.conf".

`os_sup_errortag = atom()` Defines the atom with which to tag messages received from `syslogd` before forwarding them to the error logger in the Erlang system.

SNMP MIBs

The following MIBs are defined in OS_MON:

OTP-OS-MON-MIB This MIB contains objects for instrumentation of disk, memory and cpu usage of the nodes in the system.

The MIB is stored in the `mibs` directory. It is defined in SNMPv2 SMI syntax. An SNMPv1 version of the mib is delivered in the `mibs/v1` directory.

The compiled MIB is located under `priv/mibs`, and the generated `.hrl` file under the `include` directory. To compile a MIB that IMPORTS the `OTP-OS-MON-MIB`, give the option `{il, ["os_mon/priv/mibs"]}` to the MIB compiler.

If the MIB should be used in a system, it should be loaded into an agent with a call to `os_mon_mib:init(Agent)`, where `Agent` is the Pid or registered name of an SNMP agent. Use `os_mon_mib:stop(Agent)` to unload the MIB. The implementation of this MIB uses Mnesia to store a cache with data needed. This means that Mnesia must run if this implementation of the MIB should be used. It also use functions defined for the `OTP-MIB`, thus that MIB must be loaded as well.

See Also

cpu_sup(3) [page 6], memsup(3) [page 10], disksup(3) [page 8], os_sup(3) [page 16], application(3), snmp(6)

cpu_sup (Module)

cpu_sup is part of the os_mon application and all configuration parameters are defined in the reference documentation for the os_mon application.

cpu_sup is a process which supervises the CPU load in the operating system. The load is obtained via the Solaris kernel statistics library, kstat. The same underlying mechanism is used by many other well known UNIX programs, such as rup, top and xload.

The Solaris kernel continuously maintains a large number of statistics, of which the current load values are just a few. Whenever an Erlang process requests a load measurement, cpu_sup just reads the latest statistical values.

The Solaris kernel load values are proportional to how long time a runnable UNIX process has to spend in the run queue before it is scheduled. Accordingly, higher values mean more system load. The returned value divided by 256 produces the figure displayed by rup and top. What is displayed as 2.00 in rup, is displayed as as load up to the second mark in xload.

For example, rup displays a load of 128 as 0.50, and 512 as 2.00.

If the user wants to view load values as percentages of machine capacity, then this way of measuring presents a problem, because the load values are not restricted to a fixed interval. In this case, the following simple mathematical transformation can produce the load value as a percentage:

PercentLoad = 100 * (1 - D/(D + Load))

D determines which load value should be associated with which percentage. Choosing D = 50 means that 128 is 60% load, 256 is 80%, 512 is 90%, and so on.

Another way of measuring system load is to divide the number of busy CPU cycles by the total number of CPU cycles. This method is used by some systems, including Windows NT for example, and it produces values in the 0-100 range immediately. However, this method hides the fact that a machine can be more or less saturated.

A server which receives just enough requests to never become idle would score 100% with this measurement method. If the server receives 50% more requests, it would still score 100%. With the measurement method used in this module, the load would increase from 80% to 87% when calculated with the percentage formula shown previously.

Exports

`nprocs() -> UnixProcesses`

Types:

- UnixProcesses = integer()

Returns the number of UNIX processes running on this machine. This is a crude way of measuring the system load, but it may be of interest in some cases.

`avg1() -> SystemLoad`

Types:

- SystemLoad = integer()

Returns the average system load in the last 60 seconds, as described above. 0 represents no load, 256 represents the load reported as 1.00 by `rup`.

`avg5() -> SystemLoad`

Types:

- SystemLoad = integer()

Returns the average system load from the last 300 seconds, as described above. 0 represents no load, 256 represents the load reported as 1.00 by `rup`.

`avg15() -> SystemLoad`

Types:

- SystemLoad = integer()

Returns the average system load from the last 900 seconds, as described above. 0 represents no load, 256 represents the load reported as 1.00 by `rup`.

disksup (Module)

disksup is part of the os_mon application and all configuration parameters are defined in the reference documentation for the os_mon application.

disksup is a process which supervises the available disk space in the system. Once every disk_space_check_interval minutes, the disks are checked and an alarm is generated for each disk which uses more than the disk_almost_full_threshold of available space.

On UNIX All (locally) mounted disks are checked, including the swap disk if it is present.

On WIN32 All logical drives of type "FIXED_DISK" are checked.

The disksup process defines one alarm which it sends using alarm_handler:set_alarm(Alarm). Alarm is defined as follows:

{{disk_almost_full, MountedOn}, []} This alarm is sent when a disk uses more than disk_almost_full_threshold of its available disk space, and it is cleared automatically when disksup observes that the disk space is back to normal.

Exports

get_check_interval() -> Time

Types:

- Time = integer()

Time interval, in milliseconds, which defined how often the disks are checked.

get_disk_data() -> [DiskData]

Types:

- DiskData = {Id, KByte, Capacity}
- Id = string()
- KByte = integer()
- Capacity = integer()

Gets data for the system disks or partitions. Id is a string that identifies the disk or partition. KByte is the total size of the disk or partition in kbytes. Capacity is the percentage of disk space used.

get_almost_full_threshold() -> integer()

Threshold as a percentage of the available disk space. It specifies how much disk space can be used by each disk or partition before an alarm is sent.

See Also

alarm_handler(3), os_mon(3)

memsup (Module)

memsup is part of the os_mon application and all configuration parameters are defined in the reference documentation for the os_mon application.

memsup is a process which supervises the memory usage for the system and for individual processes, as follows:

- If more than system_memory_high_watermark of available system memory is allocated, as reported by the underlying operating system, the alarm system_memory_high_watermark is set.
- If any Erlang process in the system has allocated more than process_memory_high_watermark of total system memory, the alarm process_memory_high_watermark is set.

The total system memory reported under UNIX is the number of physical pages of memory times the page size, and the available memory is the number of available physical pages times the page size. This is a reasonable measure as swapping should be avoided anyway, but the task of defining total memory and available memory is difficult because of virtual memory and swapping.

The memsup process defines two alarms which are set by the alarm_handler:set_alarm(Alarm) function. Alarm is defined as:

{system_memory_high_watermark, []}. This alarm is set when the used system memory exceeds system_memory_high_watermark of the total available memory.

{process_memory_high_watermark, Pid}. This alarm is set when an Erlang process exceeds process_memory_high_watermark of the total available memory.

These alarms are cleared automatically when the alarm cause is no longer valid.

There is also a interface to system dependent memory data, get_system_memory_data/0. The output is highly dependent on the underlying operating system and the interface is targeted primarily for systems without virtual memory (e.g. VxWorks). The output on other systems is however still valid, although sparse.

A call to get_system_memory_data/0 is more costly than a call to get_memory_data/0 as data is collected synchronously when this function is called.

Exports

`get_check_interval() -> Time`

Types:

- Time = integer()

A time interval, in milliseconds, which defines how often memory is checked. The `get_system_memory_data()` function is in no way affected by this interval.

`get_memory_data() -> MemData`

Types:

- MemData = {TotalMemorySize, AllocatedBytes, {LargestPid, PidAllocatedBytes}}
- TotalMemorySize = integer()
- AllocatedBytes = integer()
- LargestPid = pid()
- PidAllocatedBytes = integer()

Returns data about the memory in the system. `LargestPid` is the Pid of the largest Erlang process in the system. `PidAllocatedBytes` is the amount of memory the `LargestPid` has allocated.

`get_system_memory_data() -> MemDataList`

Types:

- MemDataList = [TaggedValue ...]
- TaggedValue = {Tag, Value}
- Value = integer()
- Tag = atom()

Gets system dependent memory data. The result is returned as a list containing tagged tuples, where the tag can be one of the following:

`total_memory` The total amount of memory (in bytes) available to the erlang emulator, allocated and free. May or may not be equal to the amount of memory configured in the system.

`free_memory` The amount of free memory available to the erlang emulator for allocation.

`system_total_memory` The amount of memory available to the whole operating system. This may well be equal to `total_memory` but not necessarily.

`largest_free` The size of the largest contiguous free memory block available to the erlang emulator.

`number_of_free` The number of free blocks available to the erlang system. This gives a fair indication of how fragmented the memory is.

As with `get_memory_data()`, the values on Unix-like systems indicate the amount of *physical* memory that is configured and free. The `largest_free` and `number_of_free` tags are currently only returned on a VxWorks system.

All memory sizes are presented as number of *bytes*.

`get_procmem_high_watermark() -> integer()`

Threshold as a percentage of the total available system memory. It specifies how much memory can be allocated by one Erlang process before an alarm is sent.

`get_sysmem_high_watermark() -> integer()`

Threshold as a percentage of the total available system memory. It specifies how much memory can be allocated by the system before an alarm is sent.

See Also

alarm_handler(3), os_mon(3)

nteventlog (Module)

The nteventlog module is a server which will inform your erlang application of all events written to the Windows NT event log. This is implemented with a port program that monitors the eventlog file and reacts whenever a new record is written to the log.

Your erlang application is informed of each record in the eventlog through a user supplied callback function (an “MFA”). This function can do whatever filtering and formatting is necessary and then deploy any type of logging suitable for your application. When the user supplied function returns, the log record is regarded as accepted and the port program updates it’s persistent state so that the same event will not be sent again (as long ar the server is started with the same identifier).

When the service is started, all events that arrived to the eventlog since the last accepted message (for the current identifier) are sent to the user supplied function. This can make your application aware of operating system problems that arise when your application is not running (like the problem that made it stop the last time). The interpretation of the log records is entierly up to the application.

When starting the service, a identifier is supplied, which should be reused whenever the same application (or node) wants to start the server. The identifier is the key for the persistent state telling the server which events are delivered to your application. As long as the same identifier is used, the same eventlog record will not be sent to erlang more than once (with the exception of when grave system failures arise, in which case the last records written before the failure may be sent to erlang more again after reboot).

If the event log is configured to wrap around automatically, records that has arrived to the log and been overwritten when the server was not running are lost. The server however detects this state and loses no records that are not overwritten.

Exports

```
start(Identifier,MFA) -> Result
```

Types:

- Identifier = string() | atom()
- MFA = {Mod, Func, Args}
- Mod = atom()
- Func = atom()
- Args = list()
- Result = {ok, Pid} | {error, {already_started, Pid}}
-
- LogData = {Time,Category,Facility,Severity,Message}
- Time = {MegaSecs, Secs, Microsecs}

- MegaSecs = integer()
- Secs = integer()
- Microsecs = integer()
- Category = string()
- Facility = string()
- Severity = string()
- Message = string()

This function starts the server. The supplied function should take at least one argument of type LogData, optionally followed by the arguments supplied in Args.

The `LogData` tuple contains:

1. The message `Time` is represented as by the erlang:now() bif.
2. The message `Category` which usually is one of the strings “System”, “Application” or “Security”. Note that the NT eventlog viewer has another notion of category, which in most cases is totally meaningless and therefor not imported into erlang. What this module calls a category is one of the main three types of events occuring in a normal NT system.
3. The message `Facility` is the source of the event, usually the name of the application that generated it. This could be almost any string. When matching events from certain applications, the version number of the application may have to be accounted for. What this module calls facility, the NT event viewer calls “source”.
4. The message `Severity` is one of the strings “Error”, “Warning”, “Informational”, “Audit_Success”, “Audit_Faulure” or, in case of a currently unknown Windows NT version “Severity_Unknown”.
5. The `Message` string is formatted exactly as it would be in the NT eventlog viewer. Binary data is not imported into erlang.

`start_link(Identifier,MFA) -> Result`

Types:

- Identifier = string() | atom()
- MFA = {Mod, Func, Args}
- Mod = atom()
- Func = atom()
- Args = list()
- Result = {ok, Pid} | {error, {already_started, Pid}}

Behaves as start/2 but also links the server.

`stop() -> Result`

Types:

- Result = stopped

Stops a started server, usually only used during development. The server does not have to be shut down gracefully to maintain its state.

SEE ALSO

os_sup(3) [page 16] and the Windows NT documentation.

os_sup (Module)

This module starts a server written in Erlang (and later referenced only as server), which receives messages from the Solaris operating system. The messages are tagged with an atom and subsequently forwarded to the *error_logger* in the Erlang system. If the atom is `std_error`, the messages are handled the same way as the bulk of internal error messages in the Erlang system.

This module, together with the dedicated UNIX-processes, makes a number of reconfigurations to the Solaris operating system when the service is enabled. These configurations include:

- the installation of a new configuration file for `syslogd`
- the creation of a named pipe
- the start of a port program.

As a consequence of these modifications:

1. `syslogd` writes messages of interest to the named pipe
2. the port program reads messages from the named pipe and forwards them to the server
3. the server delivers them to the error logger of the Erlang system.

When the service is disabled, the original configuration is restored.

Exports

`start() -> Result`

Types:

- Result = {ok, Pid} | {error, {already_started, Pid}}
- Pid = pid()

This function starts the server together with its dedicated UNIX processes. It returns `{ok, Pid}` if the start was successful, otherwise `{error, already_started}`.

`start_link() -> Result`

Types:

- Result = {ok, Pid} | {error, {already_started, Pid}}
- Pid = pid()

This function starts the server together with its dedicated UNIX processes. The caller is also linked to the server. It returns {ok, Pid} if the start was successful, otherwise {error, already_started} .

```
stop() -> Result
```

Types:

- Result = ok | not_started

This function stops the server and restores the original configuration of the operating system. It returns ok if successful, otherwise not_started.

Operation

1. This module is normally started by the *supervisor* and *supervisor_bridge* behaviours. Consequently, the user should not call the functions described above.
2. This module cannot be run in multiple instances on the same hardware. Special care must be taken if two or more Erlang nodes execute on the same hardware platform so that only one node uses this service *in any one instance.*
3. This module requires that a number of actions be been taken prior to starting it. These actions must be performed with *root* privileges on SunOS 5 and include change of ownership and file privileges of an executable binary file, and copying and creating a modified copy of the configuration file for the syslog-daemon *syslogd*. In addition, a the following configuration parameters must be set.
   (a) implement the server using *gen_server*.
   (b) implement protection against starting two or more instances of the service on the same hardware platform.

See also

- os_mon(3), error_logger(3), Installation Guide for your platform.
- syslogd(1M), syslog.conf(4) in the Solaris documentation.

Index

Modules are typed in *this way*.
Functions are typed in `this way`.