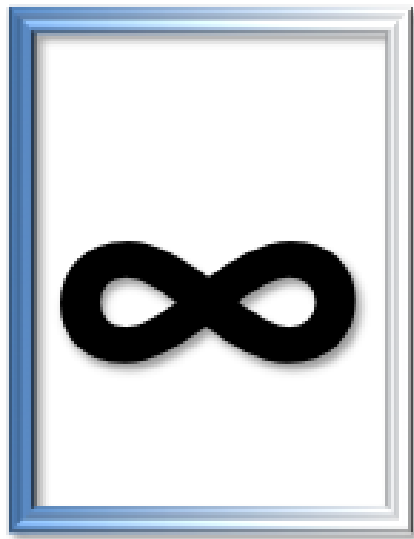


amask v1.1

AMASK: Affinity MASKs for parallel processes/threads
Document Revision 1.1
July 15, 2024



Kent Milfeld
milfeld@tacc.utexas.edu

High Performance Computing
Texas Advanced Computing Center
The University of Texas at Austin

Copyright 2024 The University of Texas at Austin.

Abstract

Amask is a set of tools for application developers and users to discover the affinity masks of application processes (MPI ranks or OpenMP threads) so they can determine where the processes can run. Amask has the following components:

- Stand-alone executables to report default masks of OpenMP, MPI or Hybrid executions in an interactive or batch environment.
- API for instrumenting applications to report affinity masks from within a program.
- Utilities: timers, set process/thread affinity, create loads (for `top` viewing)

Our intention is to create a tool that provides simple-to-understand affinity information. Bug reports and feedback on usability and improvements are welcome; send to milfeld@tacc.utexas.edu with amask in the subject line.

If you use amask, cite:

github.com/TACC/amask, “amask: Affinity Mask”, Texas Advanced Computing Center (TACC), Kent F. Milfeld. [1]

or

“amask: A tool for Evaluating Affinity Masks in Many-core Processors”, IXPUG Workshop Proceeding ISC-2017 (to be published), Frankfurt, Germany, June 22, 2017.

Contents

1	Installation	2
2	Process Affinity Masks for a Parallel Execution	3
2.1	amask Options	8
3	Using the amask Library	9
3.1	How to Print amask Reports Inside an Application	9
3.2	Useful Utilities	10
3.3	Other things you should know	11
	References	11

1 Installation

Amask is easy to build. Execute `make` to build the stand-alone executables and a library for those who want to instrument their application.

Download the github repository¹ by clicking on the “Download ZIP” file, and expanding it in a convenient location.

```
unzip amask-master.zip
```

You can also clone the git repository:

```
git clone https://github.com/TACC/amask
```

This will create a top level directory called `amask`, with subdirectories `docs` and `src`. Change directory to `amask`. Edit the `Makefile` to include an appropriate MPI compiler and OpenMP flag (the defaults are for Intel). See `INSTALL` instructions. Execute `make`.

The executables will be placed in the `amask/bin` directory and the library will be placed in `amask/lib`. Include `.../amask/bin` in your `PATH` variable.

¹<https://github.com/TACC/amask>

2 Process Affinity Masks for a Parallel Execution

Execute one of the stand-alone executables, `amask_omp`, `amask_mpi`, or `amask_hybrid` in an OpenMP, MPI or hybrid environment, respectively, to obtain the expected affinity mask for each process/thread for your program in the same environment. That is, before you execute your program, execute the appropriate stand-alone affinity program to get a listing of the masks for each OpenMP thread and/or MPI process that would occur for your program execution (provided the application doesn't adjust affinity).

In Listing 2.1 a batch job runs the `amask_mpi` executable before running an application, to observe the affinity mask for each rank that the application (`my_mpiapp`) will have. By removing the `my_mpiapp` execution line and requesting less time for the batch job, you can quickly discover the default affinity mask for the parallel environment. You can adjust affinity environment variables and quickly assess their impact on the affinity masks for your application.

Listing 2.1: Listing affinity masks for MPI environment

```

1  #!/bin/bash
2  #SLURM -n 16 -N 1
3  ...
4  #Batch Script for TACC machine
5  ...
6  mpirun ./amask_mpi      #ibrun ./amask_mpi #@TACC
7  mpirun ./my_mpiapp     #ibrun ./my_mpiapp #@TACC
8

```

If a site allows users interactive access to batch nodes (see `idev` utility @TACC), then the `amask` commands can be run interactively. Listing 2.2 shows interactive executions to discover the affinity masks that any pure OpenMP, pure MPI or a hybrid application would have for the environment. (The environment includes the OMP variables, number of MPI tasks requested, and MPI affinity environment variables set in the mpi launcher: `mpirun`, or `ibrun` at TACC.)

Listing 2.2: Listing OpenMP/MPI and Hybrid masks

```

1  export OMP_NUM_THREADS=2; amask_omp      # pure OpenMP
2
3  mpirun -np 4 amask_mpi                    # pure MPI
4
5  export OMP_NUM_THREADS=2; mpirun -np 4 amask_hybrid # Hybrid
6

```

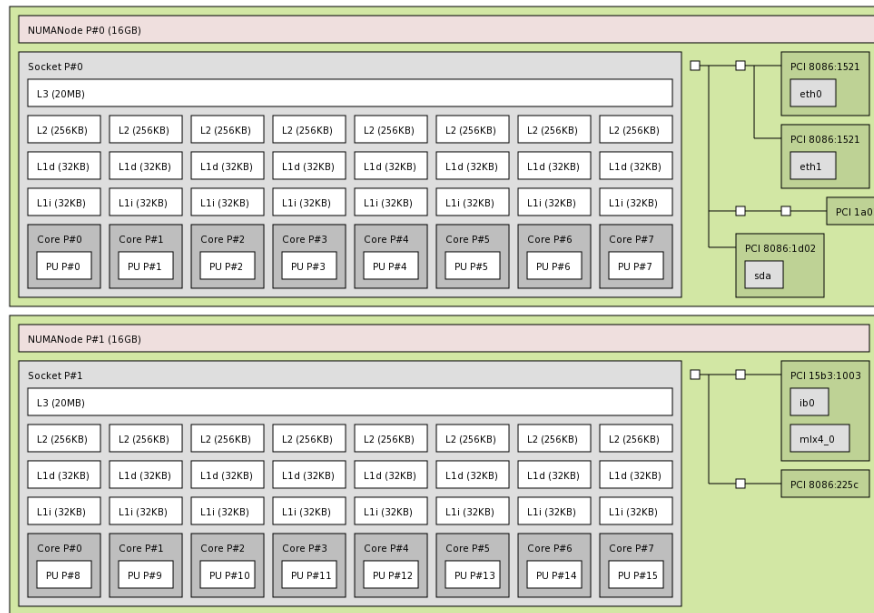


Figure 2.1: Hardware information from lstopo

How to read the output:

The output in Listing 2.3 is for an 8-rank MPI execution on a platform with 2 sockets, 8 cores/socket, no hyperthreading, and processor ids 0-7 on socket 1 and 8-15 on socket 2. One can derive these details from `lstopo`. Figure 2.1 shows an `lstopo` report for this system (Sandy Bridge compute node on the TACC Stampede machine.) The same information can be extracted from the contents of the `/proc/cpuinfo` file. The `lscpu` utility provides the number of sockets, cores and SMT threads, without `proc-id` assignments.

Listing 2.3 shows MPI processes executing on processors (cores) 0-7. The rows of the matrix are labeled by the MPI ranks and the columns represent the machine's processor-ids (Core P# in Figure 2.1) from 0 to `Nprocs-1`. For this run there are 8 rows, one for each of the 8 MPI ranks (0000 through 0007). Since there are 16 processors (`Nprocs`), there are 16 character locations in each row, representing the mask bits for processors 0 - 15. A digit in a row position means that the mask bit is set for the corresponding processor, and the MPI process rank listed at the beginning of the row can execute there. When multiple numbers are in a row (multiple bits set) it means that the rank can run (float) on any of those processors. (The - characters indicate the mask bits are not set.) The processor id number is determined by adding the digit in the row to the "group" number directly above in the header. (This quirky approach allows us to represent a processor's mask as a set of single characters (digits) for each bit, and allows one to determine the processor id (proc-id) number by a "look up"— adding the digit in the row to header group value above).

Listing 2.3: Default MPI Environment

```

1 $ # 8 tasks requested, MVAPICH2 MPI
2
3 $ mpirun amask_mpi    # non-TACC
4 $ ibrun  amask_mpi    # @TACC
5
6
7      Each row of matrix is an affinity mask.
8      A set mask bit = matrix digit + column group # in [...]
9 rank |    0    |    10    |
10 0000 0-----
11 0001 -1-----
12 0002 --2-----
13 0003 ---3-----
14 0004 ----4-----      #digit + group # in header = proc-id
15 0005 -----5-----
16 0006 -----6-----
17 0006 -----7-----

```

The output in Listing 2.4 is for the same platform as 2.3, but the affinity environment has been changed with the `tacc_affinity` script on the `ibrun` launcher line (`ibrun tacc_affinity amask_mpi`). There are other ways to change the affinity, such as using another launcher, compiling with a different mpi library, or setting MPI environment variables. In this case `tacc_affinity` positions the first 4 tasks on socket 0, and allows each rank to float on any of the processors (cores) of the socket, by setting the first 8 bits of the mask for each of the 4 ranks, 0-3. Ranks 4-7 execute on the second socket and are allowed to float across all cores (proc-ids 8-15) of the socket. (`tacc_affinity` is not part of the `amask` utilities.)

Listing 2.4: TACC Tailored Environment through `tacc_affinity`

```

1 $ # 8 tasks requested, MVAPICH2 MPI
2 $ # + tacc_affinity -- sets MPI affinity @TACC
3
4 $ ibrun tacc_affinity amask_mpi    # @TACC
5
6      Each row of matrix is an Affinity mask.
7      A set mask bit = matrix digit + column group # in [...]
8
9 rank |    0    |    10    |
10 0000 01234567-----      # -- mask to run anyplace on socket1
11 0001 01234567-----      # rank 1 can execute on any of 0-7 cores
12 0002 01234567-----      # rank 2 can execute on any of 0-7 cores
13 0003 01234567-----      # ...
14 0004 -----89012345      # -- mask to run anyplace on socket2
15 0005 -----89012345      # rank 5 can execute on any of cores 8-15
16 0006 -----89012345      # Bits are for proc-ids (cores) 8,9,10-15
17 0007 -----89012345      # Read as 8+0, 9+0, 0+10, 0+10, ... 10+15

```

Assume a system consists of 2 sockets, each with 28 cores (56 total) and each core is hyperthreaded, making 112 proc-ids. The proc-ids are numbered sequentially on each socket: even numbers on socket 0 (0,2,4,...,54) and odd numbers on socket 1 (1,2,3,...,55). On each core there are 2 Hardware (HW) threads.

The first HW threads on each core are assigned the usual proc-id with out hyperthreading. So that on socket 0, the proc-ids (for HW1) are 0,2,4,...,54, and 1,3,5,...,55 for socket 1. The 2nd HW thread are assigned by adding 56 to the non-hyperthreaded proc-id, forming the sequence 56,58,60,...,126 and 57,59,61,...,127, respectively.

The first display below shows possible occupations of 0,56, 2,58, 4,60 for threads 0, 1, and 2 (using -vs to display a single row for the mask). The second display (the default for hyperthreading, -vm for multiple lines for a mask) shows the columns as essentially core numbers, with HW threads 0,0+56, 2,2+56, and 4,4+56, respectively (as two rows with with == and - lines). The latter display (default) is much easier evaluate same-core bit settings. (With OMP_PROC_BIND=close, the affinity mask below allows a thread to occupy either HW1 or HW2.) This is shown in Listing 2.5.

Listing 2.5: Process masks for an 8-thread execution on 68-core KNL

```

1
2 $ export OMP_NUM_THREADS=3 OMP_PROC_BIND=close
3 $ amask_omp -vs
4
5     Each row is a mask for a process/thread; each char represents a proc-id.
6     A digit means the mask bit is set & process can exec on the proc-id.
7     id = digit + column group # of header between the bars (e.g. |...20...|)
8
9     *** Use -vm to generate separate row for each hardware thread.
10
11 thrd |    0    |   10   |   20   |   30   |   40   |   50   | 60...110 |
12 0000 0-----6-----...-----
13 0001 --2-----8-----...-----
14 0002 ----4-----0-----...-----
15
16 $ export OMP_NUM_THREADS=3 OMP_PROC_BIND=close
17 $ amask_omp
18
19     Hyperthreading detected: Use -vs to generate a single row for each process
20
21     Each row is a mask for a process/thread; each char represents a proc-id.
22     A digit means the mask bit is set & process can exec on the proc-id.
23     id = digit + column group # of header between the bars (e.g. |...20...|)
24     For multiple HW threads: add 56 to each additional unlabeled row.
25
26 thrd |    0    |   10   |   20   |   30   |   40   |   50   |
27 0000 0=====
28      0-----
29 0001 ==2=====
30      --2-----
31 0002 ===4=====
32      ----4-----
33  \ \ \ \

```

This listing shows core occupation in each column – something users found they would rather see than the 272-character kernel masks (of proc-ids vs. process). For a process, the sequence of rows corresponds to SMT thread-ids, 0-(n-1), where n is the number of SMT threads for each core. The first row of a process uses the “=” symbol, instead of the “-” symbol, to indicate a new process. A

digit in any SMT row can be combined with the group number in the header to obtain the core-id. To obtain the proc-id for a particular row entry, add a multiple of the number of cores (68 in this case) to the core-id, where the multiplier is the SMT thread-id number (0,1,2,3 in this case) for the core. E.g. In Listing 2.5 the proc-id for the set bit of rank 6 is 1 (the core-id) plus 2x68 (SMT thread-id 2 x number of cores). When you need to work directly with the kernel mask (e.g. to easily determine the proc-ids), force the kernel view of the mask with the `-vk` option on the command line (e.g. `amask_omp -vk`).

2.1 amask Options

There are several useful options that can be used on the `amask -omp`, `-mpi`, and `-hybrid` execution command lines. They are summarized in Table ?? .

By default, after the masks are printed to standard out, a load is created on each process for 5 seconds so that the location of the processes can be viewed with the `top` or `htop` utilities. The duration of the load can be changed with the wait option (`-w`). Include the number of seconds (integer) as an argument to the option.

By default each map is printed immediately. A 0.3sec delay can be introduced by specifying the slow argument for the printing option, `-ps`.

When HyperThreading is detected the listing defaults to a multi-line view (with a separate line for each hardware thread, HW0 and HW1, in the core). In this case, an additional constant must be added to the column value to obtain the second HW core-id. Specify `-vs` option (single) to obtain the normal Linux “kernel” view of proc-id versus process.

As described above, an argument of `s` with the listing `-v` option forces the matrix listings to be a single line representing proc-ids (without regard to hardware thread sharing); that is, it is a kernel mask.

The basic operations used for reporting masks by the `amask` executables were collected into a library, so that users could instrument there own applications to display the masks of MPI processes and OpenMP threads.

Table 2.1: amask Options for amask Executables

Purpose	option	args	long option	long args
View	<code>-v</code>	<code>s m</code>	<code>-view</code>	<code>single multiline</code>
Print speed	<code>-p</code>	<code>f s</code>	<code>-printspeed</code>	<code>fast slow</code>
Wait secs (process load)	<code>-w</code>	<code>#</code>	<code>-waitsecs</code>	<code>#</code>
help	<code>-h</code>		<code>-help</code>	

3 Using the amask Library

3.1 How to Print amask Reports Inside an Application

To report the masks from inside a program, include the amask API routine, `amask_omp()`, `amask_mpi()`, or `amask_hybrid()`, within an OpenMP parallel region, after MPI has been initialized, or within an OpenMP region of a hybrid program, respectively. These functions can be called from C, C++, and Fortran; and don't require any arguments or include files:

- `amask_omp()`
- `amask_mpi()`
- `amask_hybrid()`

However, to report masks for hybrid code, it may be necessary to initialize MPI with the `MPI_Init_thread()` routine. View the amask codes to see how easy it is to include it in your own program. Be aware, with the invocation of `MPI_Init_thread()` a thread-safe version of MPI will be used, and the MPI performance may change. The performance may be better because the newer, thread-safe library is often optimized first. The snippets below show how they are to be used:

Listing 3.1: Invoking mask report inside code

```

1  // Pure OpenMP code
2  #pragma omp parallel
3  {
4      amask_omp();
5      ...
6  }
7
8  // Pure MPI code
9  MPI_Init(NULL, NULL);
10
11     amask_mpi();
12     ...
13     MPI_Finalize();
14
15 // Hybrid code
16 MPI_Init_thread(NULL, NULL, MPI_THREAD_MULTIPLE, &provided);
17
18 #pragma omp parallel
19 {
20     amask_hybrid();
21     ...
22 }
23 ...
24 MPI_Finalize();

```

3.2 Useful Utilities

- `load_cpu_nsec(nsec)` — Puts load on process/thread for nsec seconds
- `map_to_procid(proc-id)` — Sets process/thread to execute on proc-id
- `gtod_timer()` — Easy to use Get Time of Day clock
- `tsc()` — Returns time stamp counter value

A thread or process that calls `load_cpu_nsec(int nsec)` will execute integer operations (a load) for nsec seconds. nsec must be 0 or a positive integer. Use the `cmdln_get_nsec_or_help(int *nsec, int argc, char *argv[])` function to extract an integer from the command line for nsec (e.g. `mpi amask_mpi 10`).

A thread or process that calls `map_to_procid(int proc-id)` will assign the calling thread or process to execute on proc-id, by setting the appropriate bit in the scheduling mask. For example, in a parallel region the unique thread-id returned from `omp_get_thread_num()` can be used in an arithmetic operation (linear, modulo, etc.) to create a unique proc-id to be executed on.

The function `gtod_timer()` returns a double precision number with the number of wall-clock seconds since the previous call. The first call sets the time to

zero. The function uses the Unix `gettimeofday` utility, and can be called from C/C++ and Fortran. See comments in the code for more details.

The function `tsc()` returns an 8-byte integer (Unix) time stamp count from the `rdtsc` instruction. Use this to capture the difference between the counts for determining the cost of a small set of operations (instructions, code statements).

3.3 Other things you should know

Build `amask` with the same MPI library that your application will use. On systems with multi-vendor MPI libraries, build `amask` for each vendor version you will use. With IMPI (Intel MPI) you can set `I_MPI_DEBUG=4` to separately report the mask (and other information) for each MPI process.

Bibliography

- [1] Amask is a set of executables and routines for reporting affinity information.
<https://github.com/tacc/amask>