How to install GAMMA: the Genoa Active Message MAchine

Giuseppe Ciaccio
DISI, Università di Genova
via Dodecaneso 35, I-16146 Genova, Italy
ciaccio@disi.unige.it

1 March 2007

1 Introduction

1.1 About this document

This document is about installing the Genoa Active Message MAchine (GAMMA). It provides useful information about the following topics:

- the hardware/software components needed to use GAMMA (Section 2);
- how to download and configure the GAMMA source code (Section 4);
- how to tune some GAMMA parameters (Section 5);
- how to configure the Linux 2.6 kernel for use with GAMMA (Section 6);
- how to compile and install GAMMA (Section 8);
- how to set up a convenient environment for using GAMMA (Section 9);
- how to compile a GAMMA-enhanced Linux kernel (Section 10);
- how to check whether GAMMA is installed properly or not (Section 11);
- possible causes and remedies for common troubles (Section 12).

This document is **not** about the GAMMA Application Programming Interface (API). The GAMMA API is described in the GAMMA home page on the WWW. The WWW page also contains more information concerning other aspects of GAMMA.

The WWW home page of GAMMA is at `http://www.disi.unige.it/project/gamma/`.

1.2 A sketch of the GAMMA software architecture

GAMMA is a messaging system designed for low-latency, high throughput inter-process communication across a Gigabit Ethernet LAN.

Different from many so-called “user-level” messaging system, GAMMA communications are (lightly) mediated by the Operating System (OS) of the local host computer; the GAMMA software architecture is therefore quite ”classical”, being GAMMA structured as three main components:

- a device driver, called the *GAMMA driver*, placed in the OS kernel, which implements basic communication routines;

- a set of so called “light-weight” system calls (a lower overhead implementation of the classical concept of OS system calls), called the *GAMMA system calls*, which allow a user process, running in user space, to invoke the basic communication routines, placed in kernel space, in a protected way; and
- a user programming library, called *GAMMA library*, which provides user application programmer with a convenient API for both point-to-point and (a few) collective communication routines.

There is one GAMMA driver for each supported Network Interface Card (NIC) (see Section 2). The GAMMA drivers are modified versions of the standard Linux device drivers for the same NICs.

NOTE: within the GAMMA driver, the IP functionality is blocked. But GAMMA needs IP in order to start parallel jobs (the startup is done via `rsh` or `ssh` commands). Therefore, you really need an additional LAN for the IP services. See Section 1.2.1 for details.

GAMMA provides a non-standard set of communication routines. Programmers who prefer to write parallel applications using an industry-standard API might want to install MPI atop GAMMA (see the WWW page of the MPI/GAMMA project, `http://www.disi.unige.it/project/gamma/mpigamma`).

1.2.1 GAMMA and IP: the need for an additional LAN

As pointed out before, the GAMMA driver does not support the traditional IP-based protocols because the IP functionality of the GAMMA driver is suspended. This limitation leads to better perfomance and stability. The drawback is that an additional LAN for IP is needed, because GAMMA needs the IP services when starting the parallel job. The remote process instances are indeed generated by using the `rsh` or `ssh` services, so IP must work while launching the job. Moreover, the executable file is usually made available to all nodes via a network file system (NFS, for instance), which relies upon IP as well. So an additional LAN is necessary.

1.3 How to contact the GAMMA authors

The authors of GAMMA are:

- Giovanni Chiola, `chiola@disi.unige.it`: Former project supervisor (do not contact him any more please);
- Giuseppe Ciaccio, `ciaccio@disi.unige.it`: Current maintainer; and
- other persons who contributed new GAMMA drivers; see the GAMMA web page for a complete list, or look inside source code and read about authors.

2 Requirements

2.1 User prerequisites

In order to install GAMMA successfully, you should:

- be able to configure a LAN under Linux;
- be able to configure and compile a Linux kernel; and
- have a superuser account on all the machines of your cluster.

2.2 Hardware/software requirements

The requirements to install GAMMA are:

- A pool of Personal Computers (PCs) with PCI bus.
  Supported CPUs are: Intel Pentium, AMD K6, and all superior models, including their 64-bit versions (so called "x86_64" or "amd64" architecture, not to be confused with the IA-64).

  Multi-CPU machines are allowed. It is not yet clear whether the current version of GAMMA is thread-safe from the user application standpoint, so you might have troubles in case of multi-threaded parallel jobs. Otherwise, the use of multi-CPU nodes is safe with GAMMA.

  Hyperthreading seems to cause troubles, so it should be disabled by BIOS. Multi-core CPUs have not yet been tested with GAMMA.

- Gigabit Ethernet NICs supported by GAMMA. Currently, GAMMA supports the following (families of) NICs:

  – many adapters (possibly hardwired on the motherboard) equipped with one of the Intel PRO/1000 (8254x Gigabit Ethernet) chipsets;
  – many adapters (possibly hardwired on the motherboard) equipped with one of the the Broadcom "Tigon3" Gigabit Ethernet chipsets.

  You need one NIC on each PC. Additional NICs may be present, but GAMMA will not use them. As pointed out before, a separate NIC is required to support the IP protocol, which GAMMA relies upon for spawning process instances.

- A Gigabit Ethernet switch, or a single (possibly crossover) cable for a "back-to-back" connection (two machines only).

- An additional NIC (of any kind) on each PC, and an additional hub/switch, to support IP, which GAMMA relies upon for spawning process instances (see Section 1.2.1).

- Linux kernel version 2.6.18.1

  The GAMMA distribution can be easily hacked to work with all Linux 2.6 kernels; contact the GAMMA authors in case of doubt (see Section 1.3).

- gcc 3.3 or 3.4 or 4.0.2 . Other compilers might also work fine, but we have not yet tested them.

2.3 Assumptions about file placement

For simplicity, we assume that the pathname of the Linux source tree is `/usr/src/linux/`. Other directories are allowed however, at the user choice.

3 Install GAMMA step by step

The installation procedure for GAMMA is quite complex, and assumes you have some expertise in performing various tasks (see Section 2.1). By carefully reading the following Sections, you will be able to come to success after a number of steps that must be carried out in a precise order, namely:

1. Step 1: Download and configure the GAMMA source code (Section 4)
2. Step 2 (optional): Tune some GAMMA parameters (Section 5)
3. Step 3: Configure the Linux kernel for use with GAMMA (Section 6)
4. Step 4: Set the `include/asm` symbolic link (Section 7)
5. Step 5: Compile and install GAMMA (Section 8)
6. Step 6: Set up the environment (Section 9)

7. Step 7: Compile the GAMMA-enhanced Linux kernel, and reboot the cluster (Section 10)

The installation procedure can be simplified if you are installing a new release of GAMMA having already installed a previous one. In this case, if you are re-using the same Linux kernel and no changes to the cluster environment are planned, you may skip steps 3, 4, and 6.

4 Step 1: Download and configure the GAMMA source code

4.1 Download the GAMMA distribution

Go to the GAMMA Web page, `http://www.disi.unige.it/project/gamma/`, section "Software", and click on the link pointing to the GAMMA source distribution; your browser will prompt you to download (by HTTP) a file named like "gamma-YY-MM-DD.tar.gz" (for instance: `gamma-06-02-17.tar.gz`). Get it.

For simplicity, we will call the GAMMA distribution `gamma.tar.gz`, that is, we will omit the "YY-MM-DD" part of the file name (which changes every time a new GAMMA distribution is put on the Web).

4.2 Unpack the GAMMA source code

Place `gamma.tar.gz` into your favourite working directory. Now, invoke this command:

```
tar xvzf gamma.tar.gz
```

This operation will create a subdirectory named `gamma`, containing the GAMMA source code.

4.3 Configure the GAMMA source code

Enter the subdirectory `gamma/` and run the script named `configure`. This script will ask you a few questions, concerning some configuration parameters of your local installation of GAMMA. The possible answers, along with the default answer, are indicated as well. In the remaining part of this Section, we explain the meaning of each question of the `configure` script.

```
Select the CPU architecture ["i386", "x86_64"]
```

Specify the CPU architecture of your cluster's processors. Please note that the architecture "x86_64", sometimes referred to as "amd64" (Opteron, Athlon64) has nothing to do with the IA64 architecture (Itanium). The architecture "i386", however, refers to the classical IA32 (Pentium, AMD Athlon 32 bit, etc.). In response to this question, the script creates a number of symbolic links to the appropriate files containing source code that is dependent on the CPU architecture.

```
Select the Network Interface Card (NIC) to be operated by GAMMA
```

Specify which NIC are you going to operate with GAMMA. In response to this question, the script creates a number of symbolic links to the appropriate files containing source code that is dependent on the specific NIC.

```
Enable Jumbo Frames (USE_JUMBO_FRAMES) (y/n)
```

Most Gigabit Ethernet NICs support so called *Jumbo Frames*, that is, they allow packets of greater size than the Ethernet standard. This feature, which also must be supported by the LAN switch, yields fairly greater communication throughput, and for this reason is enabled by default. You should however turn it off if the LAN switch does not support Jumbo Frames, or if you are unsure about this.

NOTE: some Gigabit Ethernet adapters do not support Jumbo Frames at all.
NOTE: Intel PRO/1000: some Intel PRO/1000 adapters (8254x chipsets) may become unreliable when Jumbo Frames are enabled, with an increased probability for the NIC to hang-up during use. The GAMMA driver for the Intel PRO/1000 can detect and solve these hang-ups, but performance may suffer; in case of frequent hang-ups, you should consider disabling this feature. See end of Section 12 for details.

```
Allow more process instances of same job on same node
(recommended with multi-CPU machines)
(MORE_INSTANCES_ON_SAME_NODE) (y/n)
```

Each parallel job is made up of a number of cooperating processes, called *process instances*. For best performance, each PC should run as many process instances as the CPUs available locally, so as to allow all instances to be simultaneously running. In case you wish to run more than one process instance on the same node (for instance, because the node is a dual-CPU), you should enable the "more process instances on same node" feature. Please note that this has *nothing to do with time-sharing the cluster among more parallel job* (a feature GAMMA always supports by default).

```
How many process instances per node
(recommended answer:  same as number of CPUs per node)
(MAX_NUM_LOCAL_INST)
```

If support for multiple instances on same node has been enabled, here it is possible to specify how many process instances are to be run on each node. It makes much sense that this be equal to the number of CPUs available per node.

```
Strip away GAMMA-related args at job launch (STRIPARGS) (y/n)
```

This option is to be considered if you want to use MPI atop GAMMA and your MPI job refuses to start (a notable example is NetPipe). GAMMA uses inline args to send some information to remote process instances when spawning them. These args can cause some picky MPI programs (e.g. NetPipe) to abort during initial phases. The `n` answer (default) is usually fine, but if your NetPipe for MPI then refuses to start, you have to change to `y` and recompile your MPI/GAMMA and Netpipe. With this option enabled, however, those MPI applications that look at the args before calling `MPI_Init()` might show startup problems; notable examples are all the NAS benchmarks. So, don't enable this option unless your own MPI programs really need it.

```
Use flow control in GAMMA barrier sync (USE_FLOWCTL_IN_SYNC) (y/n)
```

Specify whether to use flow control when performing a GAMMA barrier synchronization. Given the synchronization semantics of such a collective routine, flow control is usually unnecessary (default).

```
Print info when launching GAMMA jobs (VERBOSE) (y/n)
```

Specify whether to enable the printout of additional information at the time of launching a GAMMA job.

```
Spin-yield instead of busy-waiting on receive (SPIN_YIELD) (y/n)
```

Specify whether to use spin-yield mode instead of pure busy-waiting whenever waiting for message receive. The default answer is `y`, and is strongly recommended especially if the cluster is used in non-dedicated mode.

`Use ssh in place of rsh to spawn remote process instances (USE_SSH) (y/n)`

Specify whether to use `ssh` for spawning processes on remote nodes when launching parallel jobs. As an alternative, `rsh` can be used to this end. Default is `ssh`.

`Where is the Linux source tree installed?`

Specify the pathname of the Linux source tree on the machine where you are going to compile the GAMMA-enhanced Linux kernel (default `/usr/src/linux/`).

`Where would you like to install the GAMMA user library?`

Specify the directory where to install the GAMMA user library, once compiled (default `/usr/lib/`).

`Where would you like to install the GAMMA startup/recovery utils?`

Specify the directory where to install the GAMMA utilities for startup and recovery (see Section 9.1.3), once compiled (default `/usr/local/bin/`).

4.4 PCs with multiple NICs

Currently GAMMA is only able to drive one NIC on each PC. However, the nodes of your cluster have more than one NIC each: GAMMA needs an additional LAN for IP traffic and services (see Section 1.2.1). If the NIC used by GAMMA is not of the same "kind" as the NIC used by IP, everything is fine: the GAMMA driver will probe and operate your preferred NIC for fast communications, other Linux drivers will probe and operate the other NICs, and you may skip the remainder of this Section.

However, let us suppose you have, say, N NICs on a PC, all from the same family (that is, all of them would be driven by the same Linux driver, although they might have been sold by different vendors).

Presumably you want to use $N - 1$ of them for IP and the remaining one for GAMMA communications. Unfortunately, since all NICs are of the same kind, the GAMMA driver (like any other Linux driver) will probe all NICs and try operating them all. To avoid this, you might try and use a Linux driver together with the GAMMA driver in the same Linux kernel. However, both drivers would successfully probe all the NICs; if one of the two drivers is given precedence over the other one (as normally is when compiling device drivers in the Linux kernel), that driver will probe and operate *all* the NICs of the same kind that are detected on the PCI bus of the local host computerC, and GAMMA would not work in any case.

To allow a GAMMA driver to coexist with a Linux driver in the case of more NICs of the same kind on the same PC, the GAMMA source distribution provides slightly modified versions of Linux network device drivers for the NICs supported by GAMMA (not to be confused with the GAMMA device drivers). In the GAMMA source tree, the directory `drivers/` contains the drivers `e1000/e1000_main.c` (for Intel PRO/1000 adapters) and `tg3/tg3.c` (for Broadcom "Tigon3" adapters).

For instance, let us suppose you have N Intel PRO/1000 adapters on a PC; in order to allow GAMMA to drive one of them, and another Linux driver to drive the remaining $N - 1$ ones, do what follows:

1. Make a backup copy of the directory containing the original PRO/1000 Linux driver: `/usr/src/linux/drivers/net/e1000/`

2. In file `gamma/drivers/e1000/e1000_main.c`, set the constant `MAX_ALLOWED_INSTANCES` to $N - 1$; this will limit this driver to probe at most $N - 1$ NICs

3. Copy file `gamma/drivers/e1000/e1000_main.c` under `/usr/src/linux/drivers/net/e1000/`

4. Delete any files with suffix `.o` from directory `/usr/src/linux/drivers/net/e1000/`

Once built, your Linux kernel will use the slightly modified version of the `e1000_main.c` driver which will not conflict with the GAMMA driver.

For NICs based on the Broadcom "Tigon3" chipsets, the steps are very similar:

1. Make a backup copy of the original Linux driver: `/usr/src/linux/drivers/net/tg3.c`

2. In file `gamma/drivers/tg3/tg3.c`, set the constant `MAX_ALLOWED_INSTANCES` to $N - 1$; this will limit this driver to probe at most $N - 1$ NICs

3. Copy file `gamma/drivers/tg3/tg3.c` under `/usr/src/linux/drivers/net/`

4. Delete file `tg3.o`, if present, from directory `/usr/src/linux/drivers/net/`

The "multiple NICs of the same kind" scenario will require special attention when configuring the Linux kernel before compilation (see Section 6).

5 Step 2: Tune some GAMMA parameters

This Section is concerned about some settings of GAMMA which are not managed by the `configure` script discussed in Section 4.3).

5.1 NIC-specific settings

All GAMMA parameters whose value depends on the particular NIC in use are defined as constants in file `gamma_nic_dependent.h` (which actually is a symbolic link created by the `configure` script, see Section 4.3).

The default settings for such constants should ensure a correct behaviour of GAMMA on all platforms, so you usually do not need change anything here.

5.1.1 `GAMMA_TX_RING_SIZE` and `GAMMA_RX_RING_SIZE`

A couple of constants the value of which you might want to change are `GAMMA_TX_RING_SIZE` and `GAMMA_RX_RING_S`

Increasing the value of `GAMMA_RX_RING_SIZE` may lead to slightly better performance with GAMMA flow-controlled communication, but will consume more memory at kernel level in the OS; conversely, decreasing the value might be necessary if your PCs are short of RAM (less than 32 MBytes), but performance of GAMMA flow-controlled communication could decrease slightly (to know more about why the setting of `GAMMA_RX_RING_SIZE` may affect performance of GAMMA flow-controlled communication, see Section 5.1.2).

Increasing the value of `GAMMA_TX_RING_SIZE` may lead to better performance with GAMMA non-blocking transmission, but will consume more memory at kernel level in the OS; conversely, decreasing the value might be necessary if your PCs are short of RAM, but performance of GAMMA non-blocking transmission could suffer.

We suggest you *not* to increase the default settings, because the performance improvement obtained in this way is quite negligible. We also suggest you not to decrease the default settings unless your PCs are really short of RAM.

5.1.2 GAMMA flow control

GAMMA provides both best-effort and flow-controlled communication routines. GAMMA flow control prevents overrun at the receiver NICs from occurring, therefore avoiding packet losses at the end stations (this does not completely address the problem of packet losses in a LAN, as they might also be caused by switch congestion).

GAMMA flow control is based on *credits*. A credit is an amount of packets that a sender station is guarenteed to be able to deliver to a receiver station safely, that is, without causing a receiver ovverrun. If station S1 has a credit of C towards station S2, then S1 can safely send C packets to S2 at full speed, and C is decremented by one each time a packet is transmitted; when C is zero S1 has to stop, waiting for a "credit renewal" from S2. The larger the initial value of C, the faster the communication from S1 to S2; however, too large an initial value for C might cause an overrun to S2.

The parameter `MAX_CREDIT_SIZE` defines the maximum size of credit of each station in the cluster towards each other station. However, the parameter `LOW_WATERMARK` determines the amount of remaining credit towards a given station, below which a request for credit renewal is issued to that station. The maximum allowed values for such two parameters are subject to constraints: The latter cannot be greater than the former, and their sum cannot exceed the value $(\texttt{GAMMA\_RX\_RING\_SIZE} - 20)/(total\_num\_nodes - 1)$, which is nearly the maximum value ensuring that a receiver NIC will not overrun when simultaneously flooded by all possible senders with GAMMA flow-controlled communications.

From the above, it is clear that the largest `GAMMA_RX_RING_SIZE`, the largest the maximum value for `MAX_CREDIT_S` which translates into slightly better performance of GAMMA flow-controlled communication (as pointed out in Section 5.1.1).

Anyway, we suggest you not to change the default settings for `MAX_CREDIT_SIZE` and `LOW_WATERMARK`.

5.2 Protocol-specific settings

GAMMA parameters which affect the behaviour of GAMMA independent of the particular NIC in use are defined as constants in file `gamma_def.h`.

5.2.1 Maximum cluster size

The maximum allowed cluster size (number of PCs allowed to connect to the cluster LAN) is defined by the constant `MAX_NUM_NODES` in file `lib/gamma_userlev.h`; this is currently set to 128. Some internal data structures of GAMMA need an amount of statically allocated memory which is proportional to the maximum cluster size. Therefore, in case the PCs of your cluster are short of RAM, the default value may cause a failure during boot due to shortage of available memory. Should this ever occur, or even to prevent this from occurring, the default value for `MAX_NUM_NODES` can be set to a smaller value (e.g., 16). Of course, this imposes a limit on the effective size of your cluster. If, on the other hand, you wish to install GAMMA on a cluster with more than 128 nodes, set `MAX_NUM_NODES` to a larger value, *not exceeding 254*. Note that this defines the maximum number of nodes, not CPUs.

5.2.2 Debug kernel messages

GAMMA contains a number of statements which generate debug-oriented messages as well as warnings about communication failures. The constants `DRIVER_DEBUG_EVERYTHING` and `DRIVER_DEBUG` govern the conditional compilation of commands producing debug-oriented messages, whereas the constant `DRIVER_DEBUG_LIGHT` governs the conditional compilation of commands producing failure warnings. By default, the first two constants are "undefined" (`#undef DRIVER_DEBUG_EVERYTHING` and `#undef DRIVER_DEBUG`), whereas the third one is "defined" (`#define DRIVER_DEBUG_LIGHT`); changing the second constant to "defined" enables the compilation of code for the debug-oriented messages, and changing the first one to "defined" enables even more such messages; conversely, changing the third constant to "undefined" removes the code for emitting failure messages.
A common user will leave these settings untouched.

6 Step 3: Configure the Linux kernel for use with GAMMA

Before installing GAMMA, it is always necessary to configure the Linux 2.6 kernel in a proper way. So, after having configured GAMMA, the next step is to configure the Linux kernel.
Enter directory `/usr/src/linux/` (or whatever directory where your Linux source tree has been put), and type:

```
make menuconfig
```

or:

```
make config
```

A script (which you should be familiar with) is started, which will prompt you to answer a number of questions concerning your Linux kernel configuration.

6.0.3 General Linux kernel settings

In section `Processor type and features`, the following features *must be disabled*:

```
Preemptible Kernel

Use register arguments (EXPERIMENTAL)
```

In the same section, if your have multiple-CPU nodes (or multicore CPUs) you might want to enable the feature `Symmetric multi-processing support`, but then you *must disable* the feature:

```
SMT (Hyperthreading) scheduler support
```

Do not forget to disable Hyperthreading by BIOS as well.

Under section `Kernel hacking`, you *must disable* the feature:

```
Use 4Kb for kernel stacks instead of 8Kb
```

Under section `Device Drivers`, subsection `Network device support`, you *must disable* the feature:

```
Network console logging support
```

6.0.4 Linux kernel settings for GAMMA, Intel PRO/1000

Enter section `Device Drivers`, then enter subsection `Networking support`.
Enable features `Networking support` and `Network device support`, if not already enabled.

If you want to install the GAMMA driver for NICs based on one of the Intel 8254x Gigabit Ethernet chipsets (Intel PRO/1000 for instance), enter subsection `Ethernet (1000 Mbit)`.

Now there are two possibilities, namely:

- You have only one such NIC on each PC of your cluster. In this case, you *must disable* the feature:

`Intel(R) PRO/1000 Gigabit Ethernet support`

This will prevent the `e1000_main.c` Linux network driver from operating the NIC, letting the GAMMA driver operate safely.

- You have more than one such NIC on the PCs of your cluster. In this case, you should already have done the installation steps described in Section 4.4. Please check this out before going on.

  At this point, you *must enable* the feature `Intel(R) PRO/1000 Gigabit Ethernet support`.

  This will allow the modified version of the `e1000_main.c` Linux driver to operate on all but one of your Intel-based NICs, letting the GAMMA drive operate on one of them safely.

  If the above feature is enabled as in-kernel (you selected `Y`) then you may expect network device `eth0` to be operated by the `e1000` driver and `eth1` to be operated by GAMMA. If, however, your choice was to enable the `e1000` driver as a module (you selected `M`), then you may expect the reverse to occur, namely, `eth0` operated by GAMMA and `eth1` operated by the `e1000` driver.

6.0.5 Linux kernel settings for GAMMA, Broadcom "Tigon3" chipsets

Enter section `Device Drivers`, then enter subsection `Networking support`.
Enable features `Networking support` and `Network device support`, if not already enabled.

If you want to install the GAMMA driver for NICs based on one of the Broadcom "Tigon3" chipsets, enter subsection `Ethernet (1000 Mbit)`.

Now there are two possibilities, namely:

- You have only one such NIC on each PC of your cluster. In this case, you *must disable* the feature:

  `Broadcom Tigon3 support`

  This will prevent the `tg3.c` Linux network driver from operating the NIC, letting the GAMMA driver operate safely.

- You have more than one such NIC on the PCs of your cluster. In this case, you should already have done the installation steps described in Section 4.4. Please check this out before going on.

  At this point, you *must enable* the feature `Broadcom Tigon3 support`.

  This will allow the modified version of the `tg3.c` Linux driver to operate on all but one of your Broadcom Tigon3 NICs, letting the GAMMA drive operate on one of them safely.

  If the above feature is enabled as in-kernel (you selected `Y`) then you may expect network device `eth0` to be operated by the `e1000` driver and `eth1` to be operated by GAMMA. If, however, your choice was to enable the `e1000` driver as a module (you selected `M`), then you may expect the reverse to occur, namely, `eth0` operated by GAMMA and `eth1` operated by the `e1000` driver.

7 Step 4: Set the `include/asm` symbolic link

After having configured the Linux kernel, the subsequent step is to make sure that the symbolic links `include/asm` exists.

If not, create it by simply starting the kernel compilation (command "make") and interrupting after a handful of seconds. The kernel compile procedure creates that symlink before actually compiling, and for now we just need the symlink to be created (we shall compile the kernel later).

8 Step 5: Compile and install GAMMA

After configuration and tuning, and after the Linux kernel has been configured correctly, the GAMMA source code can be compiled and installed. Here, by "installation" we mean the integration of the GAMMA code into the Linux kernel; therefore, it will be necessary to recompile the Linux kernel after having installed GAMMA (see Section 10).
Enter the subdirectory `gamma/` and compile GAMMA by invoking

```
make
```

The following warning might show up during compilation:

```
Warning:  indirect lcall without '*'
```

It depends on the version of assembler being used by the compiler. You can ignore it.
Next, install GAMMA by invoking

```
make install
```

9 Step 6: Set up the environment

9.1 Cluster-specific environment

9.1.1 Hyperthreading must be disabled by BIOS

Do not forget to disable Hyperthreading by BIOS. This feature may show instabilities with GAMMA, which are not yet investigated.

9.1.2 GAMMA configuration file

GAMMA needs a configuration file called `/etc/gamma.conf`, containing global information about the cluster. More precisely, `/etc/gamma.conf` must contain the mapping between each PC hostname and the MAC address of the NIC that will run the GAMMA communications. Here is a sample `gamma.conf` file for a cluster with four machines:

```
felix 0x00 0x00 0xf8 0x1b 0x39 0xbb
orion 0x00 0x00 0xf8 0x1b 0x3a 0x0c
aries 0x00 0x00 0xf8 0x1b 0x37 0x17
gemini 0x00 0x00 0xf8 0x1b 0x4b 0x43
```

The *same* configuration file must be placed in all PCs of the cluster.

The host names to be put in the file can be obtained by running the command `hostname` on each PC. MAC addresses can be obtained with the command `ifconfig`. Comments and blank lines are *not* allowed in the file.

9.1.3 GAMMA startup and recovery utilities

When you have configured the GAMMA source code, the `configure` script asked you the following question:

```
Where would you like to install the GAMMA startup/recovery utils?
```

In what follows, let us suppose the default answer, namely `/usr/local/bin/`, was selected (the following discussion, however, can be adapted to an arbitrary pathname).

Enter directory `/usr/local/bin/`, and find the following GAMMA utilities in the form of executable files:

- GAMMA utility for network startup (`gammagetconfig`)
- temptative recovery from network lockups (`gammareset, gammaresetall, gammaresetvm`)
- launching GAMMA applications written in FORTRAN (`gammarun, gammarunvm`)
- adjusting the size of GAMMA flow-control credits (`gammacredit, gammacreditall`)
- adjusting the duration of timeout for missing packets(`gammamaxpolls, gammamaxpollsall`).

You must copy all of them on the other PCs in your cluster, again under `/usr/local/bin/`.

Purpose and use of such utilities are described in Section 11.

9.2 User-specific environment

The launch of a GAMMA parallel application on the cluster is accomplished by remote shell ("ssh"). Therefore, you have to properly set some configuration files in all the PCs of your cluster, in order to be able to run commands via "ssh" as a user without being asked the user password. Of course you must hold a user account on each PC.

The shell variable PWD must be present in the user environment. For instance, users who prefer the bash shell should add the following statement to their `.bash_profile` file:

```
export PWD
```

To launch a GAMMA parallel job, one copy of the executable must be present on each PC in the cluster and all the copies must have the same absolute pathname and filename. The obvious way to enforce this, is to have one single copy of the executable on a directory of one PC and let all the other PCs to share that directory remotely via NFS; to this end, the cluster administrator might have to act upon files `/etc/fstab` and `/etc/export` on some or all the PCs. Please note that in a large cluster this might be impractical (when starting the parallel job, a huge contention would put the NFS server into troubles).

10 Step 7: Compile the GAMMA-enhanced Linux kernel, and reboot the cluster

After the installation of GAMMA, it is always necessary to recompile the Linux 2.6 kernel.

Enter directory `/usr/src/linux/`, then compile the Linux kernel as usual (”make”). Place the new Linux kernel on all PCs of your cluster, run `lilo` on all of them if needed, then reboot everything.

11 GAMMA at work

11.1 Checking for successful NIC probing

After the boot, check the boot messages on all cluster machines, using the command `dmesg`. If you see a message like:

```
GAMMA Genoa Active Message MAchine (Linux 2.6)
```

followed by a calendar date and other information, then you are sure that the GAMMA driver was successful in detecting and probing your preferred NIC.

11.2 Starting up the network after a boot

To start up the GAMMA network after a boot, first configure the network device as usual (that is, using the Linux `ifconfig` utility). In case you have configured GAMMA to use Jumbo Frames, read the MTU size from the GAMMA boot message in `dmesg`, and use that value to configure the network device.

After configured the network device, run the GAMMA `gammagetconfig` utility (see Section 9.1.3), by invoking it with no arguments.

Usually, the LAN configuration is done at boot time by a boot script (somewhere under `/etc/rc.d/`); you might consider the possibility of editing that script (or other boot-time scripts) in your machines, in order to add the invocation of `gammagetconfig` so as to configure the GAMMA network automatically after a boot.

11.3 Running a GAMMA parallel job written in C

As already pointed out in Section 9.2, in order to launch a GAMMA application, one copy of the executable must be present on each PC in the cluster and all the copies must have the same absolute file name. The best way to enforce that, is to have one single copy of the executable on a directory of one PC and let all the other PCs to share that directory remotely via NFS. Also, make sure the shell variable PWD is exported in the user environment.

To launch a GAMMA parallel job, simply type the name of the executable followed by the in-line arguments required by the program itself. For instance, to launch a GAMMA parallel program whose name is `ping_pong` and requiring one numeric in-line argument with value 5, simply type

```
ping_pong 5
```

Note that no indication is given as for the degree of parallelism. With GAMMA, the number of process instances to spawn must be explicitly set up by the program itself once launched (of course the program can always get the desired degree of parallelism from the user through an in-line argument).

In the above example, no indication is given concerning which PCs are to host the process instances of the parallel job. By default, the first instance (with instance number 0) always runs on the local machine (the one the job has been launched from); the names of other machines needed for other process instances are taken from file `/etc/gamma.conf`, by reading the file circularly starting from the name just after the one of the local machine. For instance, let us suppose file `/etc/gamma.conf` is like this:

```
felix 0x00 0x00 0xf8 0x1b 0x39 0xbb
orion 0x00 0x00 0xf8 0x1b 0x3a 0x0c
aries 0x00 0x00 0xf8 0x1b 0x37 0x17
gemini 0x00 0x00 0xf8 0x1b 0x4b 0x43
```

If a GAMMA job is started by machine `gemini` and generates three process instances, then instance 0 runs on `gemini`, instance 1 runs on `felix`, and instance 2 runs on `orion`. If however the GAMMA jobs generates five instances, then instances 0, 1 and 2 are as before, instance 3 runs on `aries`, and instance 4 on `gemini`; in this case, one machine (`gemini`) hosts two process instances (0 and 4).

The user is allowed to designate explicitly which machines are to host the process instances of a job. For instance, let us suppose the user wishes to run a GAMMA job called `barrier` with four process instances, using the two machines `gemini` and `orion` so that instances 0 and 1 run on `gemini` and instances 2 and 3 on `orion`. To this end, the user must first create a *machine file* like this:

```
gemini
orion
orion
```

then must log on `gemini` and launch the job from that machine. If the name of the above machine file is `mach`, to launch the job the user should type:

```
barrier 4 -machinefile mach
```

The GAMMA runtime support will then spawn the process instances so that the first one (instance 0) runs on the local machine (`gemini`), whereas the other three instances run on the machines listed in the machine file, in the same order.

11.4 Sample GAMMA parallel C programs

Under directory `gamma/apps/pingpong/` you can find a few simple GAMMA parallel jobs for latency and bandwidth measurements, which can also be used for testing the functionality of GAMMA. A Makefile is provided to compile them.

The most important of such programs is `ping_pong.c`. The executable `ping_pong` performs a number of message round-trips in order to measure the average GAMMA communication delay between two machines. To run it, you must simply invoke `ping_pong` followed by an argument, namely, the size in bytes of the message to be exchanged. If size is zero, then the program will output the one-way latency time (half the round-trip time) in μsec; otherwise, it will output the one-way throughput in MByte/s, computed as the message size divided by half the average round-trip time. The source code of `ping_pong` contains a number of options, in the form of C constants that can be "defined" or "undefined" at will; for instance, you can choose the specific GAMMA communication routine to be used for the test (best-effort, flow-controlled, etc.) by acting over constants `USE_2P`, `USE_ISEND`, `USE_FLOWCTL`.

Another simple but interesting GAMMA parallel program is `barrier.c`, under `gamma/apps/barr/` (there is also a Makefile for building the executable). Once compiled, the executable `barrier` takes one argument, namely, a number `N`, and measures the average barrier synchronization time among `N` processes in the cluster. `N` is not allowed to exceed the total no. of nodes in the cluster multiplied by the maximum number of process instances allowed per node.

For other GAMMA applications, explore the directory `gamma/apps/`. Please note, most of these sample programs are very old and not maintained at all.

11.5 Running GAMMA parallel jobs written in FORTRAN

Under `gamma/apps/pingpong/fortran/` you can find a sample "ping-pong" GAMMA parallel program written in FORTRAN, plus a Makefile to compile it (which however uses the `f2c` translator, instead of the `f77` compiler). Basically, this program is there only to test if it can compile and run.

To launch it, you must use the utility script `gammarun` (see Section 9.1.3). Type:

gammarun 2 ping_pong

where the "2" specifies the number of computing nodes involved in the application run.

11.6 Restarting the network after a crash

GAMMA is still far from being a perfectly stable communication system. If something goes wrong during a run of a GAMMA parallel application, and your network appears to be no longer working, you might try the following recovery procedure which works in most (but not all) cases.

Let us suppose the device operated by the GAMMA driver is seen as `/dev/eth1` on all the PCs. Then, on each PC invoke:

```
gammaresetvm
gammagetconfig
```

As a better alternative, on just one of the PCs (at your choice) you might invoke the script `gammaresetall` (see Section 9.1.3).

Even better, run the script `gammacreditall -r` (see Section 11.7).

11.7 Adjusting the size of GAMMA flow control credits

In some cases, it might be helpful to act upon the size of the credits in the GAMMA credit-based flow control (see Sections refflowctl.

To adjust the credit size, you can run the utility `gammacredit` (see Section 9.1.3) on each PC in the cluster, followed by `gammaresetall` (see above). As an alternative, on just one of the PCs (at your choice) you might invoke the script `gammacreditall` (see Section 9.1.3 again).
For example, by typing:

```
gammacreditall 12
```

the credit size is set to 12 packets on each machine.
To restore the credit size to the default value, type:

```
gammacreditall -r
```

To read the current credit size, simply type:

```
gammacreditall
```

11.8 Adjusting the timeout for missing packets in the GAMMA retransmission algorithm

GAMMA implements mechanisms to detect and possibly retransmit missing packets. To allow detecting packet losses, each GAMMA packet travelling from node A to node B is tagged by a unique id number $Id(A, B)$. The current value of $Id(A, B)$ is incremented by one at each packet transmision from a to B; thus, any node can detect packet losses by simply checking the sequences of id numbers tagging its incoming packets.

This mechanism works enough in most cases; there is an exception, however, namely: if a packet get lost, and no more packets are transmitted, the missing id number cannot be detected. It is therefore necessary to run a mechanism based on a *timeout* at the receiver side, in order to catch all possible packet losses.

In some cases, it might be helpful to act upon the duration of the timeout. Too short a timeout, indeed, might force unneeded packet retransmissions; on the countrary, too long a timeout turns out into slower recovery from packet losses.

To tune the duration of the timeout, you can run the utility `gammamaxpolls` (see Section 9.1.3) on each PC in the cluster, followed by `gammaresetall` (see above). As an alternative, on just one of the PCs (at your choice) you might invoke the script `gammamaxpollsall` (see Section 9.1.3 again).
For example, by typing:

```
gammamaxpollsall 5000
```

the timeout is set to approx. 1 sec on each machine.
To restore the timeout to the default value, type:

```
gammamaxpollsall -r
```

To read the current setting for the timeout, simply type:

```
gammamaxpollsall
```

12 Quick troubleshooting

This is a list of common troubles that you might encounter when using GAMMA. For each symptom there is a temptative diagnosis and one or more suggestions for actions that could help solve the problem.

This Section is not to be inteded as complete. Future versions of this document might hopefully add more paragraphs to this Section, also with the help of GAMMA users.

Symptom The boot of a GAMMA-equipped PC fails, showing error messages seemingly related to an "out of memory" scenario.

Suggestions The PC is short of RAM, and a file system check has been issued during boot. The file system check consumes a lot of kernel memory and apparently does not release it when finished. In this case, simply reboot the machine in the hope the file system check be skipped this time. However, if the problem shows up even if no file system check has been carried out, then you have to decrease the maximum allowed cluster size; read Section 5.2.1 to know more about this.

Symptom You try to launch a GAMMA application but nothing seems to happen.

Suggestions It may depend on many causes. If GAMMA was configured to use Jumbo Frames, check that the MTU size for the ethernet device was set correctly (Section 11.2, and that the NICs and the LAN do support Jumbo Frames. Check whether the GAMMA driver was able to detect and probe the NIC (Section 11.1). Check whether the GAMMA configuration file `/etc/gamma.conf` (Section 9.1.2) is correct, is present on all PCs and is the same on all PCs. Check whether the executable file of the GAMMA application is seen on all the involved PCs with the same absolute pathname. In doubt, recompile the GAMMA library with "verbose" mode turned on (Section 4.3), then recompile the application and try again.

Symptom You try to launch a GAMMA application and the following message appears on the standard output:

```
gamma_init():  Could not read /etc/gamma.conf
```

Suggestions The GAMMA configuration file `/etc/gamma.conf` (Section 9.1.2) could not be read on one or more PCs in the cluster. Check whether `/etc/gamma.conf` is correct, is present on all PCs and is the same on all PCs.

Symptom You try to launch a GAMMA application and the following message appears on the standard output:

```
gamma_init():  could not create a virtual GAMMA
```

Suggestions GAMMA failed to create some local runtime data structures. The most likely reason for this, is that many GAMMA applications were badly killed before this attempt. Invoke the utility `gammacreditall -r` (Section 11.7) on all the PCs in the cluster, then try again. If GAMMA was configured to use Jumbo Frames, check that the MTU size for the ethernet device was set correctly (Section 11.2, and that the NICs and the LAN do support Jumbo Frames.

Symptom You try to launch a GAMMA application and the following message appears on the standard output:

```
gamma_init():  registration failed
```

Suggestions GAMMA failed to create some runtime data structures across the cluster. This might indicate a communication problem on the LAN. Check whether the GAMMA driver was able to detect and probe the NIC (Section 11.1). If GAMMA was configured to use Jumbo Frames, check that the MTU size for the ethernet device was set correctly (Section 11.2, and that the NICs and the LAN do support Jumbo Frames Check whether the GAMMA configuration file `/etc/gamma.conf` (Section 9.1.2) is correct, is present on all PCs and is the same on all PCs.

Symptom You try to launch a GAMMA application and the following message (or a very similar one) appears on the standard output:

```
gamma_init():  launch on node 3 failed!
```

Suggestions GAMMA failed to spawn one or more remote processes of the parallel application. Check whether you can run commands on remote nodes via `ssh` (Section 9.2). Check whether the shell variable PWD is exported in the user environment (Section 9.2). Check whether the executable file is seen on all the involved PCs with the same absolute pathname (Section 9.2).

Symptom You try to launch a GAMMA application and the following message (or a very similar one) appears on the standard output:

```
bash:  /root/pingpong/ping_pong:  No such file or directory
```

Suggestions Check whether the executable file is seen on all the involved PCs with the same absolute pathname (Section 9.2).

Symptom You try to launch a GAMMA application and the following message (or a very similar one) appears on the standard output:

```
bash:  syntax error near unexpected token ‘(null)/ping_pong’
```

```
bash: -c: line 1: '(null)/ping_pong 0 -GAMMAHOME (null)/ -GAMMA 1'
```

Suggestions Check whether the shell variable PWD is exported in the user environment (Section 9.2).

Symptom You try to launch a GAMMA application from a PC in the cluster and the PC itself suddenly crashes.

Suggestions This is a rare consequence of a (not yet fully understood) bug affecting the initial phases of a GAMMA job launch. The only solution is to reset the crashed PCs, run the `gammacreditall -r` script, and try launching again.

Symptom A GAMMA parallel program stops or freezes during run, and the `dmesg` command on all or some processing nodes shows messages similar to these ones:

```
sys_gamma_send(): par_pid 1, out_port 1, prog_no 3: Failure

sys_gamma_send(): par_pid 1, out_port 1, prog_no 3: Fail after 10 trials
```

Suggestions This usually indicates a LAN hardware problem; check your LAN hardware (especially the cables). It might also indicate a very congested LAN, where a transmitting NIC could not send packets for too long a time.

Symptom During execution of a GAMMA parallel jobs, `dmesg` command on all or some processing nodes shows the following message, repeated a number of times:

```
sys_gamma_poll(): no pkts for a long time. Testing frame seq # and credits...
```

Suggestions This might indicate three very different conditions, namely:

- The duration of the retransmission timeout (Section 11.8) is too short. This forces numerous yet unneeded checks for missing packets. Use the `gammamaxpollsall` utility (see Section 11.8) to decrease the value of the timeout.
- The running parallel job congests the LAN switch, and therefore a lot of packets get lost. This might result into slower completion of the job. You probably need a different, more performing but possibly more expensive interconnect; as an alternative, you might try to modify the parallel program in order to enforce a better communication pattern among cooperating processes of the parallel job.
- A process instance `A` of your parallel job expects messages from some other process instance `B`, but `B` is not transmitting (because of reasons depending on your application). After waiting for a while, `A` will probe the network for possible loss of messages. You can ignore this normal behaviour, unless you see that your job does not make progress for too long a time.

Symptom The GAMMA test program `ping_pong` works with messages of size below 1500 bytes, while does not work with larger messages.

Suggestions You configured GAMMA to use Jumbo Frames (see Section 4.3), but your NIC and/or switch does not support them, or the MTU size for the network device is not set to the proper value. In the former case, configure GAMMA again and make sure the Jumbo Frames feature is disabled. In any case, check out and set the correct MTU size for the network device (see Section 11.2).

Symptom Your parallel job experiences occasional slow-downs or temporary hang-ups, and the following message frequently appears in the output of the `dmesg` command at some nodes of your cluster:

```
tx_flush_ring():  NIC restarted
```

Suggestions You configured GAMMA to use Jumbo Frames with the Intel PRO/1000 NICs (see Section 4.3). These NICs may become unreliable when Jumbo Frames are enabled. The probability of hang-up increases if your parallel job is such that each node sends messages to numerous other nodes, or the MTU size is greatly larger than the standard 1500 bytes. If you experience this kind of problem, configure GAMMA again and disable the Jumbo Frames feature. Also, make sure to set the standard MTU size for the network device (see Section 11.2).