Configuring Amanda for Parallel Backups

By Gregory Grant
Revision 01

Abstract

In today's computing environment running backups in parallel can often provide several benefits. These include the reduction of the backup window and increased throughput to the backup media. Understanding how to control your backup system's capability to run parallel backups is important if you want to optimize your backup environment.

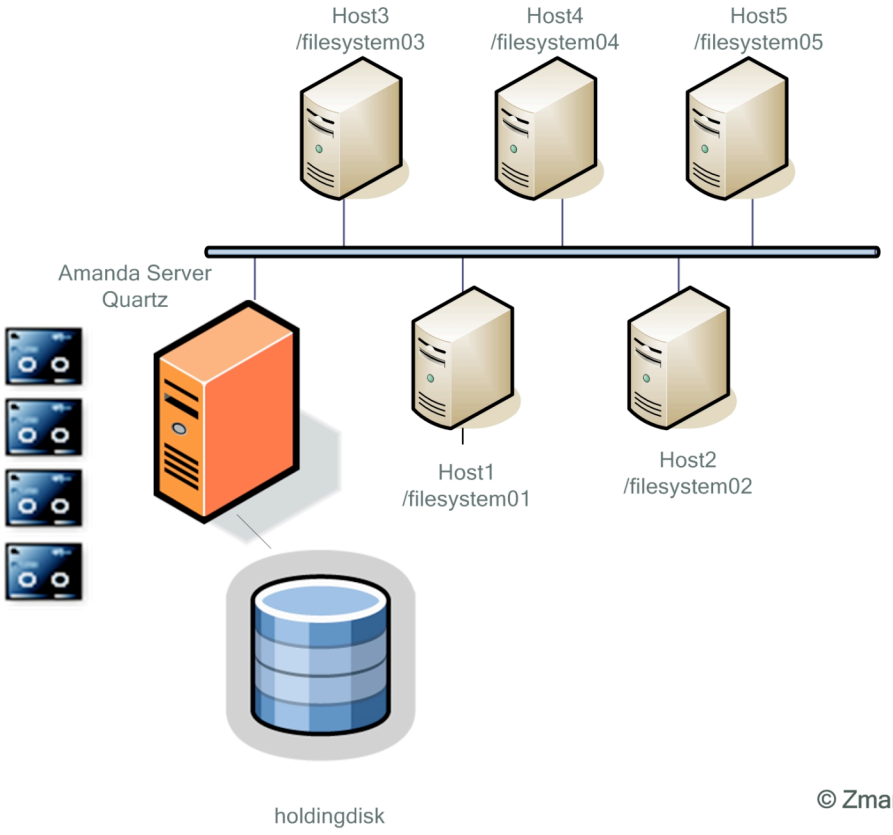
Amanda allows you to back up client systems in parallel. This short article describes some of the parameters that control parallel backups, and how they might be used to optimize your backups.

Audience

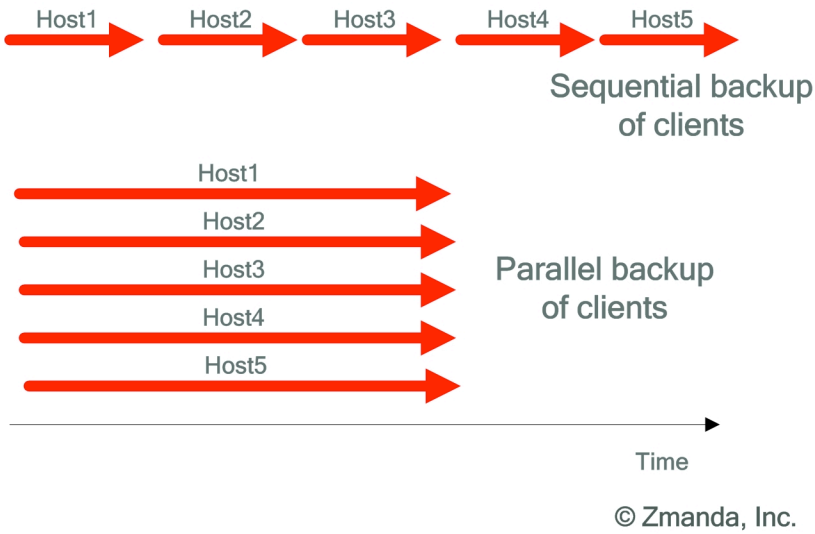
This paper is intended for people that are relatively new to Amanda planning and configuration. However, we assume you are familiar with the Amanda configuration files amanda.conf and the format of the disklist file. If you are unfamiliar with either of these you can refer to the Amanda man pages amanda.conf and amanda. (see http://wiki.zmanda.com/index.php/Man_pages). We also assume that you have been introduced to the concept of holding disks.

Introduction to Amanda Parallel Backups

Below is a typical Amanda backup environment. In this environment the Amanda server has five clients to back up, each with a single file system. If you have adequate resources (memory, CPU, disk space, network bandwidth, etc.) you can configure Amanda to run these backups in parallel.



The illustration below shows the potential benefit of running the backups in parallel – the total time for the backup can be less in a parallel backup.



Amanda uses several parameters and settings to determine the number of parallel backups. In summary they are:

- The amount of space available on the holding disk(s).
- The amount of estimated network bandwidth that will be consumed by each backup.
- The *Parallel Backups* setting in Zmanda Management Console, which is the *inparallel* setting in the amanda.conf file. This parameter limits the total number of parallel backups dispatched by the Amanda server. The default is 10.
- The *Parallel Backups (clients)* setting in Zmanda Management Console, which is The *maxdumps* setting for a given *dumptype*. This parameter limits the number of parallel backups dispatched for a single client.

We will examine each one separately, and include several examples of how they interact.

How Holding Disk Space Affects Simultaneous Backups

Holding disks are used by Amanda to temporarily store backups before they are written to media. Holding disks provide many benefits, including caching backups before they are written to tape media. Another benefit is the ability to run backups in parallel. Amanda requires holding disks to run backups in parallel.

During a backup, Amanda will examine how much holding disk space is available. Since Amanda estimates the size of each individual backup, Amanda can estimate how many backups, and which ones, can be dispatched simultaneously.

Let's use our example. Amanda has five file systems to back up. When Amanda runs the backup it determines the size of today's backups for each file system:

host1	/filesystem001	100 GB
host2	/filesystem002	5 GB
host3	/filesystem003	40 GB
host4	/filesystem004	25 GB
host5	/filesystem005	30 GB
Total		200 GB

Not Using a Holding Disk

Under some circumstances the holding disk will not be used at all. For instance, for the above example this can happen if the holding disk is not sized appropriately. As you can see, the smallest backup for today is 5 GB (/filesystem002). If the holding disk has less than 5 GB of available space, then Amanda has no choice but to skip the use of the holding disk. This will cause the backups to flow directly to the backup media. It further means that only **one** backup will be done at a time.

There are other circumstances under which a holding disk will not be used:

- Specifying *holdingdisk never* for the *dumptype* being used (see amanda.conf).
- When there are no holding disks defined in the amanda.conf file.

Using a Holding Disk

As stated before, if you want simultaneous backups then you must have one or more holding disks. But how does the available space on the holding disks affect the number of simultaneous backups that can be performed?

Amanda knows the amount of holding disk space available. Amanda also knows the size of each backup to be performed. Amanda combines these two to determine how many backups can be run in parallel.

For example, let us assume that we have a single **holding disk with 75 GB** of available space. A second assumption is that we are dumping the smallest backups first (you can control this behavior). This means that given our original five backups the following will be dispatched by Amanda and dump to the holding disk:

host2	/filesystem002	5 GB
host4	/filesystem004	25 GB
host5	/filesystem005	30 GB
Total		60 GB

What about the remaining file systems? Since we have defined the backups to proceed from the smallest to the largest, /filesystem03 (40 GB) will be written to tape while the three file systems above are written to the holding disk.

Amanda knows that /filesystem001 will never fit on the holding disk. Since it is the largest backup, it will be dispatched last, and will go straight to tape.

By increasing the holding disk to 100 GB we could achieve simultaneous backups of all 5 file systems: four would be dumped to the holding disk, while the largest is dumped to tape. The four being written to the holding disk will be flushed to the tape when the largest dump has completed.

If we increase the holding disk size to 200 GB, then there would be enough room to dump all 5 file systems to the holding disk at the same time.

Network Bandwidth Estimated Usage

Amanda will attempt to limit the number of backups based on estimated network bandwidth usage. If you are not careful, it is possible to configure Amanda with sufficient holding disk space and still not get parallel backups.

While Amanda uses estimated bandwidth usage in its dispatching algorithm, Amanda does not attempt to dynamically monitor and throttle bandwidth usage once a backup is started. Once a backup starts, it will use as much of the network as it can, leaving throttling and traffic shaping up to the operating system and network hardware.

When planning the backup, in addition to the size of each backup Amanda estimates the bandwidth each backup will need. Amanda does a very good job generating this estimate, since this estimate is based on historical backup data. For example, for our five file systems Amanda might determine the following bandwidth estimates:

host1	/filesystem001	400 Kbps
host2	/filesystem002	400 Kbps
host3	/filesystem003	500 Kbps
host4	/filesystem004	500 Kbps
host5	/filesystem005	600 Kbps
Total		2,400 Kbps

Specifying the Maximum Bandwidth to be used

Amanda allows you to define the total bandwidth you wish to use; this is the *netusage* parameter in the amanda.conf file. When Amanda dispatches backups in a backup set, it limits the dumps it starts to approximately the *netusage* value.

For example, if we define *netusage* to be 2,000 Kbps then Amanda will not be able to dispatch all of the above backups in parallel. Assuming we have adequate holding disk space, the following backups will be scheduled (remember we are dumping smaller backups first):

host2	/filesystem002	400 Kbps
host4	/filesystem004	500 Kbps
host5	/filesystem005	600 Kbps
host3	/filesystem003	500 Kbps
Total		2,000 Kbps

Specifying The Maximum Bandwidth per Interface

Amanda also has the concept of maximum bandwidth usage per "interface". These interfaces are given a name, but do not necessarily map to an actual network interface name defined to the operating system. Associated with the interface is the maximum bandwidth Amanda will dispatch to that interface. Each interface is defined in the amanda.conf file. An example might be:

```
define interface fromchicago {
      comment "Allow 500 Kbps to be dispatched"
      use 500 Kbps
}

define interface local {
      comment "Allow 1000 Kbps to be dispatched"
      use 1000 Kbps
}
```

The above interface definitions are straightforward – when the first interface is used Amanda will allow up to 500 kilobytes per second of backup data to be dispatched; the second will allow up to 1000 kilobyes per second to be dispatched.

But how does Amanda know which interface a backup will use?

Defining which Backups use which Interface

For each entry in the disklist file the interface to use is defined. This is the last parameter on each line of the disklist file. If one is not specified then the interface "local" is used.

```
host1 /filesystem01     dumptype001 local
host2 /filesystem02     dumptype001 fromchicago
host3 /filesystem03     dumptype001 local
host4 /filesystem04     dumptype001 fromchicago
host5 /filesystem05     dumptype001 local
```

How does this alter which backups Amanda will dispatch? Based on the order of dumping (smallest to largest) Amanda would like to dispatch the backups for /filesystem02 and /filesystem04 first. However, both of these are defined to use *fromchicago*, which specifies to only dispatch up to 500 Kbps. /filesystem02 is estimated to run at 400 Kbps, so it can be dispatched. However, /filesystem04 cannot be dispatched, since it is estimated to need 500 Kbps.

The same process is repeated for the remaining backups in the list, which are specified to use the *local* interface. Up to 1500 Kbps of backups will be dispatched.

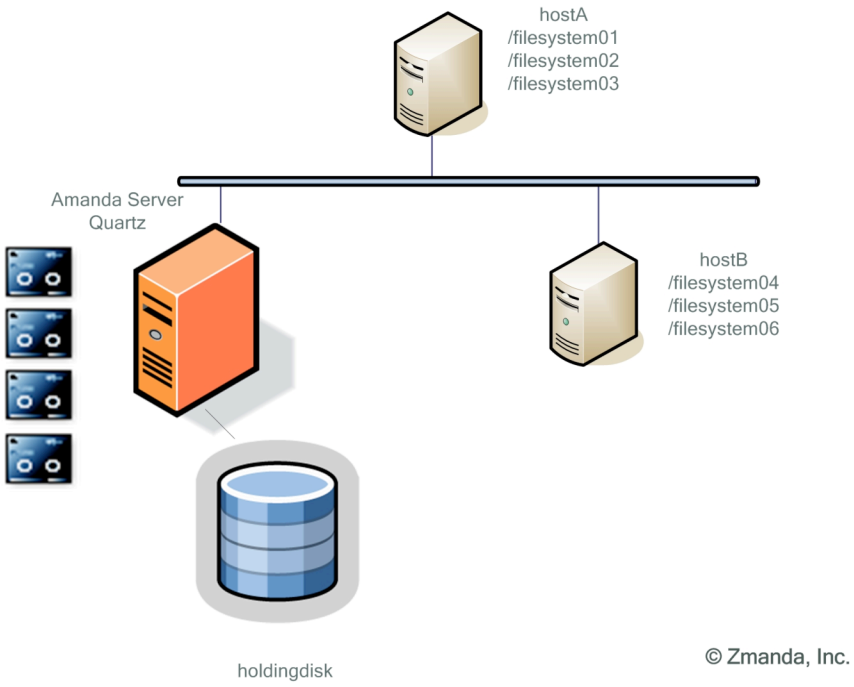
The Parallel Backups Parameter

Zmanda Management Console (ZMC) has a *Parallel Backups* parameter, which specifies the maximum number of backups that Amanda will attempt to run in parallel. It is stored in the amanda.conf file as the *inparallel* parameter. This value will not override the holding disk and network bandwidth considerations. What this means is that even if your *inparallel* value is 10 (which is the default), if there is no holding disk space, or if your *netusage* parameter isn't set appropriately, then you will not get parallel backups.

The Parallel Backups (Clients) Parameter

Zmanda Management Console (ZMC) has a *Parallel Backups (Clients)* parameter, which specifies the maximum number of simultaneous backups to run from a single client. It is stored in the amanda.conf file as the *maxdumps* parameter. The default value is "1". This means that if there are multiple entries for a given host in the disklist file, that only one at a time will be backed up.

The following example shows six file systems from two different hosts:



Let's assume that you have enough holding disk and network bandwidth settings to back up all six file systems in parallel; in order to make this happen you will need to change the *Parallel Backups (Clients)* parameter in Zmanda Management Console to "3", since there are three file systems on each client. Note that running these dumps in parallel is

good if the three file systems are on different spindles. If the three file systems are on the same spindle, then dump performance can suffer due to disk head thrashing.

Proving Benefits with a Test

But do parallel backups actually improve backup performance? To illustrate the benefits of parallel backups a test was conducted. The test environment consists of:

- Two clients with a total of 1.2 GB of data.
- Each client has one file system to back up.
- Each client is a virtual machine hosted on a single physical machine.
- The server is a second physical host.
- A DLT7000 tape drive in a tape library.
- 100 mbps network speed.
- Full backups are run for each test.
- The Amanda command line utility amstatus is used to verify whether or not parallel backups are taking place.

In the first test no holding disk was used, thus no parallel backups were performed. It took 498 seconds to complete the backup of both clients.

In the second test a holding disk was used. Further, the network bandwidth settings were set to allow for parallel backups. It took only 352 seconds to complete the backup of both clients, which is a **savings of approximately 30%**. An interesting note is the amount of time it took the clients to dump to the holding disk: 184 seconds. From the client system's perspective this represents a significant amount of savings, since once this step is complete no further resource demands are made of the client systems.

Conclusion

As you can see, Amanda gives you great flexibility, allowing you to specify how parallel backups are to be controlled:

- Holding disks
- Network bandwidth
- The global *inparallel* parameter
- The *maxdumps* parameter

By utilizing these parameters you can optimize your backups to perform the backups in the shortest possible backup window. You should carefully monitor your backups to make sure you are getting the amount of parallel backups that you need and desire.