CommunityDNS Slave Server

Introduction

The CommunityDNS Server Slave is a bootable CD that allows anybody who wishes to join the CommunityDNS effort in raising the standard of DNS delivery over the entire Internet.

Once a server have been deployed and connected into the CommunityDNS service infrastructure, it will automatically carry all zones that are part of the CommunityDNS effort and will join the Anycast cloud that ensures the unrivalled reliability in the delivery of the zones.

About the software

The CommunityDNS Server Slave is shipped as an entire package including the operating system, DNS software and all the necessary configuration to join the Anycast cloud. During the installation process the operator will be prompted for the IP Address, Subnet Mask and Default Gateway / default route for the server and nothing else.

The Operating system is based on the 2.6 Linux Kernel but is a custom distribution designed specifically for the task. It therefore only carries software that is necessary for the task, reducing the security risk carrying unnecessary packages might represent.

Both the operating system and the data are stored on the disk encrypted using the AES algorithm and a 256-bit key. This ensures the maximum protection for your zone information and the zone information that is carried on behalf of other members of the CommunityDNS project.

However, the disk does not carry the key. The key will be sent to it over a SSL connection once the machine has been inspected and verified to be genuine. Only than can it unlock the root partition and then the data partition. Both the root and data partitions use different keys where the key to the data partition is unique to each machine. If the hard disk is booted in a different machine a new key will be required and the data will be have to be reloaded from the original source.

All data sent to the server is done via an SSL VPN tunnel using a Private Certificate Authority to validate the upper level server that is providing the information, this also uses 256-bit AES.

Updates to the zone information are always sent in IXFR style, i.e. only changes are sent, and data integrity is ensured using a database style journal transaction mechanism. Updates are sent by streaming these journal transactions to ensure as little latency occurs between different nodes receiving the updates.

If a box has been off line it will automatically request updates from the last known transaction it has stored. This ensures that it can recover after a failure and that it doesn’t miss any updates. The server software supports connecting to a number of upper level journal servers and it will automatically fail over between them.

Installing the CommunityDNS Server Software

The software for the CommunityDNS Server Slave is distributed in the form of a bootable CD. To run the server, download the CD ISO image from the CommunityDNS web site (http://www.CommunityDNS.net/iso/), burn it onto a blank CD, boot the CD in the target machine and the CD will pretty much do the rest – note, this include re-formatting the hard disk. This CD has not been tested with any virtual machine architecture so our current recommendation is "don't" or "use is at your own risk" – which ever you prefer.

Note, before you install the CD you should decide what IP Address, Subnet Mask and Default Route / Gateway you wish to use for the server as these can only be assigned at install time and are then fixed.

You may install the software onto a hard disk in a different machine from the one you wish to run the server on and then simply transfer the disk to run. However, the authorisation process includes both the hardware and the IP Address, so you should only request that a machine is authorised to receive data once the disk is running in the machine you wish to keep it in.

Every time you boot the server, it will only boot the Kernel then will wait for authorisation before loading the operating system. At this point you will see a clock at the bottom of the screen. You can test if the server is working by seeing if it will respond to a "ping". At this point you will need to contact the CommunityDNS team to authorise the server to continue. At this point it would be safe to power the system off without shutting it down.

Authorisation requires that the server can be contacted on TCP port 22 (SSH) – this may require changes to your firewall. Once it has been authorised, except in the case of emergency, port 22 (SSH) could then be closed off.

Note, if at this point, you try and connect to the server using a standard ssh client, you will see the clock stop, but your ssh client will do nothing. As soon as you break the TCP connection you will see an SSL error message on the console of the server and it will reboot itself.

One of the last messages you will see in a normal boot sequence is "Entering run level 2". At this point it will try and establish the SSL Tunnel to bring in the journal stream. Once the VPN is established it will start up the DNS server and you will see the message "Entering run level 4". The server is now up and running normally, BGP will start and the server should re-join the anycast cloud.

What ports will the server use for “normal” operation

Server’s Port	Remote Port	Protocol	Use
22 (SSH)	Client Port	TCP	Authorisation at start-up
Client Port	840	UDP	Encrypted Journal Stream
53 (DNS)	Client Port	UDP	Serving DNS
53 (DNS)	Client Port	TCP	Serving DNS
22 (SSH)	Client Port	TCP	Emergency Maintenance
179 (BGP)	Client Port	TCP	BGP communication with router
Client Port	179 (BGP)	TCP	BGP communication with router
799	799	UDP	Lan broadcast for clustering
153	Client port	UDP	Server to server comms for clustering

By “client port” we mean any port in the range 1024 to 65,536. The CommunityDNS Slave Server will always use ports in the range 20,000 to 40,000.

Configuring BGP

A critical part of Anycast is that, while DNS service is available from the server, the Anycast subnet is announced via BGP and the route is withdrawn if service is lost. This ensures that queries are only routed to that Anycast node while it is active.

For this purpose, the server will announce the subnet 194.0.1.0/24 from our AS number – 42909. You will need to configure one of your routers to accept and propagate this route. We will then configure the BGP on the server to connect to the appropriate router. We will therefore need to know the IP Address of the router you wish us to configure our BGP to connect to and if you wish us to include any additional BGP options, for example "multi-hop".

If you cannot take our BGP announcement from the server, and propagate it, then you will not be able to host a CommunityDNS server, as this is such a critical part of running an Anycast service.

Our Security model

It is vitally important to us that all zone data is kept confidential. We therefore use a number of mechanisms to ensure this.

1. All servers must be authorised every time they reboot. If we suspect that someone is trying to break our security model we simply stop authorising the server to boot. Authorisation is a manual process that includes an inspection of the server.

2. All new slaves must be manually configured into our system to be allowed to receive the journal stream. We can withdraw this authorisation at any time.

3. All our servers are issued with a unique certificate by our private certificate authority. The slaves use this to verify they are talking to a bona-fide information provider.

4. If you wish to advertise our AS number via BGP we have to authorise that first, ensuring servers can't joint our anycast cloud without permission.

5. Updates to zones are transferred as updates only, not the entire zone, and encrypted with 256-bit AES

6. All data and programs stored on the disk are encrypted with 256-bit AES. The operating system and data partitions use different keys.

7. The operating system shipped on the CD is shipped pre-encrypted. No keys to access the operating system, or data, are ever stored either on the hard disk or the CD.

Hardware Specification

The only requirement for running this software is a PC compatible computer with 8Gb RAM, a compatible Ethernet controller and a single hard disk. The software is designed to run on pretty much any PC compatible hardware. We have tested it with both Intel and AMD based systems. Generally we'd recommend using a SATA disk, as they are a good balance between speed and price.

You will need an optical drive to install the software, but it is truly plug and play, so you can install the software in one machine then transfer just the hard disk to another machine for running it in.

Qualified Servers

Although the software is designed to run on almost any PC hardware that uses supported components, we have also qualified certain branded server platforms.

Sun Fire X2200 M2 Server

http://www.sun.com/servers/x64/x2200/

Although the X2200/M2 was the server in the range that we qualified we would expect any of the servers in the Sun Fire x64 range to work. The advantage of the X2200 is that it has 16 RAM slots enabling you to fit the lower cost 1Gb DIMMs.

With two dual-core 1.8GHz Operton processors and 667Mhz RAM we could achieve 82,000 queries per second on this platform.

Dell Poweredge 860 and Poweredge 2950

Again, although these were the specific models we qualified, we would expect pretty much all the server's in Dell's range to work. In fact even the entry level 2400 workstation will run the software correctly.

Supermicro

All Supermicro servers based on the Intel chipset with Intel Ethernet controllers should work.

CPU

Minimum specification: Single-Core 2.8Ghz Athlon / Operton / Single-Core 3Ghz Xeon
Preferred specification: Dual-Core 2.8Ghz Athlon / Operton / Dual Core 3Ghz Xeon

Up to 32 CPU cores are supported. Query response rate is roughly linear based on the number of cores and clock speed. Retaining a high response rate is not only important to maintaining a satisfactory service, but also to resist a Denial of Service attack.

The Slave platform is based on the Linux kernel. To get the widest range of compatibility while taking advantage of some of the features of more modern processors the processor option "`586/K5/5x86/6x86/6x86MX`" was chosen to compile the kernel. It is known this is **not** compatible with the Via C3 processor, but that processor is probably too slow for this application. This is what the Linux kernel documentation says about this processor option: -

> `Select this for a 586 or 686 series processor such as the AMD K5, the Cyrix 5x86, 6x86 and 6x86MX. This choice does not assume the RDTSC (Read Time Stamp Counter) instruction.`

RAM

Minimum: 8Gb

Hard Disk

Minimum Size: 80Gb, Preferred Size: 160Gb

PATA, Supported Chipsets: AMD / nVidia , Intel PIIX & ICH, ServerWorks OSB4/CSB5/CSB6, SiS5513 and VIA82CXXX

SATA, Support Chipsets: Intel ESB ICH PIIX3 PIIX4 MPIIX, Silicon Images, VIA – future release will include SiS 964/965/966/180 and Silicon Image 3124/3132

SCSI, Supported Chipsets: Adaptec AIC7xxx U160, Adaptec AIC79xx U320, Adaptec AIC94xx SAS/SATA

Ethernet Interface

The server only requires a single Ethernet interface.

10/100, Supported Chipsets: 3cr990 series "Typhoon", Broadcom 4400, Intel(R) PRO/100+, RealTek RTL-8139 C+, RealTek RTL-8129/8130/8139, SiS 900/7016, VIA Rhine

10/100/1000, Supported Chipsets: D-Link DL2000, Intel(R) PRO/1000, Realtek 8169, SiS190/SiS191, SysKonnect, VIA Velocity, Broadcom Tigon3, Broadcom NetXtremeII

Optical Drive

Installation of the operating system is from a downloaded ISO image. This will require an optical drive to boot from. SCSI, IDE and USB CD or DVD drives should all be supported.

Installing the Server – Step by Step

Checking the BIOS

If the server you are planning to use has an "on-board" or "integrated" graphics chip then the chances are that the graphics chip will use a proportion of the computer's main memory for graphics. However, this server package does not require a graphical user interface so any memory assigned to the graphics chip will be unavailable to the Operating System and will therefore be wasted.

So it is our recommendation that you go into the configuration options of the BIOS and reduce the amount of memory assigned for graphics use to the minimum amount allowable, typically 4Mb.

You will usually find this on a BIOS page entitled "Integrated Peripherals". However, BIOSes vary considerable from one vendor to another and from one version to another, so you may simply have to hunt around to find the option.

Downloading and Burning the CD Image

Once you have chosen a suitable server, and have checked the BIOS configuration, you are ready to install the CommunityDNS Slave Server software. The software is distributed as an ISO image that can be downloaded from our web site http://www.communitydns.eu/iso/

It will be called "encdns_<ver1>_<ver2>.iso" where "<ver1>" is the version of the installation package and "<ver2>" is the version of the actual server software. Once you have downloaded the ISO image you will need to burn it onto a blank CD.

An ISO Image is a block-by-block copy of a CD. We have to distribute the software in this format to make the CD bootable. Therefore simply copying the image file onto a CD (e.g. using copy and paste) will not work correctly. You need to use a program that can copy the image back onto a blank CD block-by-block.

Windows XP is capable of doing this, but Microsoft does not provide an interface to the function, so you may have to download an install a third party Power Toy. More information can be found here http://www.petri.co.il/how_to_write_iso_files_to_cd.htm .

It is also common that software shipped with CD-RW / DVD-RW drives (e.g. Nero) has the capability to write ISO images block-by-block. If you copy the ISO image to your desktop and Windows assigns it an icon, showing it recognises the format, then it is likely you already have some software that can write an ISO image to CD.

We have tested the ISO image successfully with both CD-R and CD-RW disks, but not with writeable DVDs, so we would not recommend their use. Although once written, the CD should boot correctly in a DVD drive.

Booting the CD in the Target Machine

Once you have written the ISO image to a blank CD next you need to boot the CD in the target machine.

When the CD first boots the install package will run a number of checks.

- If you have less then 8Gb of RAM then it will issue a warning, but allow you to continue with the installation. Note, even if you have installed 8Gb it may not all be available to the Operating System if some has been assigned to an "on-board" or "integrated" graphics chip.
- If it can not find a supported IDE (PATA), SATA or SCSI hard disk it will warn you and halt the install
- If it cannot find a supported Ethernet Controller it will warn you and halt the installation.
- If it finds that the target disk is already partitioned, it verify you are happy to erase it, then it will ask you to boot the CD a second time to proceed with the actual install. This prevents issues that can arise (especially with IDE disks) associated with disk geometry alignment.

Towards the end of the boot, it will ask you for the `IP Address`, `Netmask` and `default Gateway / Router` for this server. Note, once installed, the only way to change these would be to re-install from the CD. This is because of the way the encryption works.

Once you have entered these, the install will be finished, the server will eject the CD and prompt you to reboot from the hard disk.

Entering the three network configuration items are the only prompts you should get during the install, unless it has been necessary to blank the partition table.

Once the server has booted from the hard disk it should get as far as a ticking clock at which point the boot process will stop and wait for us to authorise it to continue. This will happen **every time** the server is rebooted. You should therefore avoid rebooting the server unless absolutely necessary.

Check the Operating System can access all the RAM

Once the Operating System has booted, you can check that it has detected all the memory and that the graphics card is not using it.

When the boot process stops at the ticking clock, press Ctrl-S to pause the clock, then press Shift-PageUp to scroll back through the boot log looking for a like that starts "Memory:" which should look something like this: -

```
Memory: 8311388k/8912896k available (2614k kernel code, 76072k reserved,
1197k data, 252k init, 7471040k highmem)
```

8Gb is exactly 8,388,608 Kb so the 8,311,388 Kb reported here is fine.

FAQ

1. If we host a server why aren't we allowed login access?

The servers carry the zone information from a number of different sources and we have contractual obligations to ensure this data is only made available in the manner prescribed. Although the data on the disk is encrypted there are various known Side Channel attacks to AES that require the ability to run software on the server, preventing login access to the server mitigates against this kind of attack.

Your zone data is as important and confidential as everybody else's. If you wish your zone to be kept confidential, then it is only reasonable to expect the same from everybody else.

Generally, during normal operation of the server, nobody will be allowed login access to the box – that is, the password on all login accounts will be disabled, preventing all login access. In the event of an emergency, a password can be restored.

All the necessary configuration of the server is done at a central point, allowing zones to be added or dropped and extra IP Address to be added into the Anycast cloud as other parties join the CommunityDNS effort.

2. If I need to shut the server down or reboot it, how do I do that?

If you press the power button then the server will automatically shut itself down and power off. If you attach a keyboard, you can use Alt-Ctrl-Del to make the server shutdown cleanly and reboot.

Generally it is not safe, or recommended, to switch the server off without shutting it down.

3. How do I change the server's IP Address?

Because of the security model we use, it is currently not possible to change the server's IP Address except by re-installing from the original CD.

4. How does the CommunityDNS effort assist in preventing Denial of Service attacks?

There are two main ways in which the CommuntyDNS service mitigates against D/DoS attacks.

Firstly, it uses Anycast. Anycast allows the same IP Address to be available at multiple locations on the Internet. The more locations, the more difficult it will be to bring down the service. BGP (the Internet's core routing protocol) then ensures that a query to a CommunityDNS server is routed to the closest available server. The software is configured to ensure that if the DNS server goes down, the BGP route is withdrawn and so that server will no longer receive queries.

By “hiding” the server behind these anycast addresses, it is much harder for an attacker to attack any one individual machine. It is therefore vital that you keep the server’s private IP Address confidential.

Secondly, the software is one of the fastest DNS servers around today. On hardware costing less than US$2000 it can handle 100,000 queries per second on a database of over 100,000,000 names. By having a large number of servers that can deliver that kind of performance its possible to simply “out gun” a D/DoS attacker.

The Anycast cloud can currently handle 25 billion queries per day and will shortly rise to 33.5 billion.

The more people who join the CommunityDNS effort, and install a server locally, the more we can ensure that we eradicate the ability of a D/DoS attack to bring down your zone. By having a server within your national address space you are helping protect your country from D/DoS attacks and providing faster responses to the zone that are participating in the CommunityDNS effort.

We are constantly putting research and development effort into increasing the performance of the software. By doing so we are constantly increasing the resistance the service has to D/DoS attacks.

5. Does the Slave support Load Balancing and / or Fail Over (Clustering)?

Yes, if you install two (or more) servers and plug them into the same network segment, they will automatically detect each other, form a cluster, decide who will be cluster controller and share the query load.

If any server goes down the cluster will automatically adjust, even if that server was the cluster controller. You can add and remove nodes from the cluster without any further configuration and up to 250 servers can join the one cluster.

You cannot have more than one cluster on the same network segment.

While any server is running you will see “heart-beat” packets broadcast on UDP port 799 for the purpose of knowing which servers in the cluster are still running.

6. What happens if we’re hosting a server and it goes down?

If any server in the Anycast cloud goes down, BGP should ensure that all queries are diverted to the next nearest server. This usually happens within about 30 seconds. Once a server comes back up, it will usually take between one and two minutes before it starts receiving queries again.

Note, every time the server is rebooted it will require authorisation before it can start back up. It is therefore highly preferable to avoid rebooting the server as much as possible. Further more, if a BGP route keeps going up and down too much upper level routers may decide the route is “flapping” and suspend it for a period of time while it is allowed to settle.

All servers in the CommunityDNS Cluster are monitored 24 x 7 x 365 by our staff in the UK, US and Japan using automated monitoring tools, which send out SMS alerts when critical systems fail.

7. I don't believe anyone can serve that much DNS at that speed from those spec servers

Well, it's not exactly a question, but it is frequently brought up. The only answer is if you don't believe it; try it for yourself.

Actual measured figures are that on a database of 500,000,000 names a Supermicro server with two dual-core 3.2Ghz Xeon processors achieved just over 100,000 queries per second and a Sun X2200 with two dual core 1.8Ghz Operton processors achieved just over 85,000 queries per second.

When we increased the database from 300,000,000 names to 500,000,000 names there was no degradation in performance, which is exactly in line with what the mathematics had predicted. We would predict an 11.5% decrease in performance if the database were doubled from 500 million names to 1,000 million names.

With a two disk stripped RAID on Ultra-320 SCSI disks we can achieve a peak update rate of 60,000 transactions per second; on a single SATA disk that falls to about 10,000.

To rebuild a database of 100,000,000 names from the journal stream takes about 1 hour.

We continue to invest in research and development to ensure that we are always getting more and more out of the hardware. Looking for the bottlenecks in the system and trying to find ways round them. We already have a road map for a development program that we are confident will continue to significantly increase performance while reducing memory and bandwidth requirements.

8. Is my Zone data secure?

Every effort has been made to ensure that the zone data is kept as secure as possible.

- All zone data is transferred to the CommunityDNS Slave Server over an SSL tunnel using 256-bit AES encryption.
- All zone data is stored on the disk using 256-bit AES.
- The data is stored using a different key from the operating system.
- The data's key will be unique to the server it is stored on.
- The zone is only ever transferred to the slave in its entirety once.
- The zone data is stored in fragments making it difficult to re-produce the entire original zone file.

Consider the contrast between this and using a plain text AXFR to a "bind" server that will then store the zone as a text file.

Where as the CommunityDNS server does not support out-bound AXFR or IXFR. There is simply no code in there to support it.