

Unicode and the Implications of Its Implementation

Contents

1. Introduction	1	5. How Snapshots Interact With Unicode	3
2. What is Unicode?	1	6. Backing up Unicode-Enabled Snap Servers.....	3
3. Converting to Unicode	1	6.1 Backing up Unicode-enabled Windows clients.....	3
3.1 Create a disaster recovery image of your system and volume settings.....	2	6.2 Backing up UNIX clients utilizing a specific codepage	3
3.2 Backup all data on your system to tape or to another system	2	6.3 Backing up Unicode-enabled MacOS clients.....	3
3.3 Convert to Unicode	2	7. Other Third-Party Applications.....	3
3.4 Create a new disaster recovery image	2	7.1 Applications bundled with GuardianOS	3
3.5 Backup with Unicode-enabled backup applications.....	2	7.1.1 NetVault BakBone.....	4
4. Unicode and Protocol Interaction.....	2	7.1.2 NDMP	4
4.1 SMB (or CIFS).....	2	7.1.3 Computer Associates eTrust Antivirus.....	4
4.2 AFP	2	7.1.4 Snap Enterprise Data Replicator (Snap EDR)	4
4.3 NFS	2	7.1.5 Server-to-Server Replication (S2Sv2)	4
4.4 FTP protocols.....	3	8.0 Unicode and Expansion Arrays	5
4.5 HTTP	3	9.0 Summary and Other Issues to Consider.....	5

1. Introduction

This whitepaper is intended to detail how the latest GuardianOS operates when Unicode is enabled. Unicode was introduced as part of v4.0.228 of GuardianOS. The following topics will be covered:

- What is Unicode?
- Converting to Unicode
- Unicode and protocol interaction
- How snapshots interact with Unicode
- Other third-party applications
- Backing up Unicode-enabled Snap Servers
- Unicode and expansion arrays

2. What is Unicode?

GuardianOS 4.0.228 and higher now supports Unicode. Unicode enables a single software product to be targeted across multiple platforms, languages, and countries without re-engineering the existing environment. It allows data to be transported through many different systems while preserving the file and directory names. More information is available at <http://www.unicode.org/>.

There are many different Unicode codepages. GuardianOS v4.0.228 utilizes the UTF-8 codepage. Enabling Unicode on GuardianOS will not affect the contents of any files, only the file and directory **names**.

Caution: Once Unicode has been enabled on a GuardianOS Snap Server, it is NOT possible to disable Unicode. Enabling Unicode could adversely affect any third-party applications that do not support the UTF-8 Unicode codepage. Strongly consider that all third-party applications accessing the Snap Server after Unicode is enabled have the capability to support Unicode. If an application does not support Unicode, the integrity of the file and directory names may be severely compromised for file and directory names that contain extended Unicode characters.

3. Converting to Unicode

Prior to enabling Unicode on your Snap Server, verify the Snap Server is running GuardianOS v4.0.228 or later. Since the Unicode feature was not available prior to v4.0.228, you must update the system before continuing.

The following sections will describe the steps necessary to convert your GuardianOS Snap Server to Unicode.

3.1 Create a disaster recovery image of your system and volume settings

Before converting your system to Unicode, configure all system and volume settings, and then create a disaster recovery image (DRImage) of your system and volume(s). This is to ensure all your settings and data are saved should something unforeseen happen during the Unicode conversion process. See your GuardianOS User Guide or the online help for more information on how to create a DRImage.

3.2 Back up all data on your system to tape or to another system

Back up your system as you normally would. For more information about backing up your server, please see your backup applications documentation.

3.3 Convert to Unicode

Within the Web Browser Administration Tool, navigate to the **Server > Unicode** screen and enable Unicode. Remember, once it has been enabled on your Snap Server you cannot disable Unicode. Once any Snap Server or expansion array has been converted to UTF-8, there is no way to “un-convert” the file system or disable Unicode without destroying all of your data. Be sure your settings and volumes have been saved to an off-server location before enabling Unicode.

Caution: Do not convert to Unicode if your volume is full. Unicode requires space on the volume for a reference file. If the volume is full, Unicode will not convert the system properly. A good rule of thumb would be to make sure that a minimum of 10% of the total volume space is available on every volume prior to enabling Unicode.

3.4 Create a new disaster recovery image

Once your system has been converted to Unicode, make a new DRImage. The procedures are the same as described above.

3.5 Back up with Unicode-enabled backup applications

Back up your system with a Unicode compliant backup application. Please see the following section **Backing up Unicode-enabled Snap Servers** for more information.

4. Unicode and Protocol Interaction

On a Unicode-enabled Snap Server, extended characters in file and directory names are encoded on the Snap Server file system using UTF-8, a method of representing all Unicode characters. However, network protocols and clients vary in their support of Unicode and UTF-8, which has ramifications in the way they interact with one another when sharing files with extended characters in filenames. The following describes how different protocols interact with extended characters.

4.1 SMB (or CIFS)

Most Windows clients and the SMB protocol support the majority of Unicode characters. Therefore, in general, all

characters written by Windows clients will be properly retained and visible to other Windows clients and Unicode-compliant protocols. However, if there are characters on the file system that cannot be mapped to the Windows Unicode encoding method (UCS-2) or written as invalid UTF-8, an escape sequence will display in the filename of the file being read.

Escape sequences begin with `{!^`. The two characters following the escape sequence are the hexadecimal value of the characters in the filename; for example, you might see `{!^AB` in a filename. Windows clients can edit such files, and the names will be retained in their original form when written back to the file system.

4.2 AFP

MacOSX and higher use the same method to represent Unicode characters as the Snap Server: UTF-8. Information written to the server from MacOSX or higher will be encoded with UTF-8 and should be viewed correctly from the MacOS UI. However, similarly to SMB clients, characters in filenames that are incompatible with UTF-8 will be returned with an escape sequence.

Escape sequences begin with `{!^`. The two characters following the escape sequence are the hexadecimal value of the characters in the filename; for example, you might see `{!^AB` in a filename. MacOSX clients can edit such files, and the names will be retained in their original form when written back to the file system. MacOS 9 and lower are not Unicode-compliant, and use the MacRoman codepage to represent extended characters. AFP translates MacRoman into UTF-8 when writing to Snap Servers. Any extended characters on the file system that cannot be translated to MacRoman will also be returned with an escape sequence.

4.3 NFS

Depending on how a NFS client is configured, the NFS protocol may not be Unicode-compliant or Unicode-aware. There is no means for the Snap Server to determine what codepage is being used by the NFS client to represent extended characters. Currently, the codepages most commonly used in Linux environments are: 8859-1, 8859-15, and EUC-JP. The Snap Server then must make an assumption to enable it to translate to and from UTF-8 on the file system. Therefore, when in Unicode mode, you must configure the Snap Server's NFS protocol for the codepage being used by NFS clients. Any extended characters on the file system that cannot be translated using the configured NFS codepage will be returned to the NFS client with an escape sequence.

Escape sequences begin with `{!^`. The two characters following the escape sequence are the hexadecimal value of the characters in the filename; for example, you might see `{!^AB` in a filename.

It is important to note that if using any other Unicode codepage other than the three codepages supported under NFS, only other

NFS clients configured the same way will be able to access and view the file and directory names correctly. Access from CIFS or any other file access protocol is not supported if using any codepages under NFS other than those listed above.

4.4 FTP protocols

FTP only supports ASCII characters by specification. Some clients bend the specification to allow extended characters, but there is no standard means of representing them. Therefore, no translation is performed on extended characters for FTP clients — all filenames are written to and read from the file system as a “bag-of-bytes”.

This has two ramifications: extended characters written to the file system by other protocols will be visible to FTP clients as garbled characters; and FTP clients are able to write invalid UTF-8 characters to the file system. For the latter case, when other protocols encounter invalid UTF-8 characters on the file system, the invalid characters will be returned in an escape sequence.

Escape sequences begin with `{!^`. The two characters following the escape sequence are the hexadecimal value of the characters in the filename; for example, you might see `{!^AB` in a filename.

4.5 HTTP

HTTP integrates easily with Unicode and the Snap Server. If invalid UTF-8 characters are encountered on the file system, the characters will be returned with an escape sequence.

Escape sequences begin with `{!^`. The two characters following the escape sequence are the hexadecimal value of the characters in the filename; for example, you might see `{!^AB` in a filename.

5. How Snapshots Interact with Unicode

Snapshots taken before the Snap Server was converted to Unicode are not compatible with the Snap Server once it has become Unicode enabled. It is not recommended that a pre-Unicode snapshot be used to restore a post-Unicode server.

Tip: It is recommended, if you have snapshots on your Snap Server that were created prior to Unicode conversion, you delete all snapshots once Unicode has been enabled on the Snap Server.

6. Backing up Unicode-Enabled Snap Servers

Backing up a Unicode-enabled Snap Server requires a third-party backup application capable of supporting the UTF-8 Unicode codepage. Although a backup application that does not support the UTF-8 Unicode codepage may appear to back up your data properly, upon any restoration, file, and directory names may not be preserved properly.

For backup applications that do not explicitly support the UTF-8 Unicode codepage, in many cases the incompatibilities exist only in the User Interface display. The backup application may be able to properly back up and restore all data, and even, potentially, preserve all file and directory names properly. Be sure to contact

your backup vendor to completely understand the extent of support for the UTF-8 Unicode codepage.

6.1 Backing up Unicode-enabled Windows clients

When backing up a Unicode-enabled Windows client, connect and back up using SMB. It is recommended that you use Veritas Backup Exec to back up Unicode-enabled Windows clients, but any Unicode-compliant backup application will also work.

6.2 Backing up UNIX clients utilizing a specific codepage

UNIX clients run one of three codepages: 8859-1 (US), 8859-15 (Europe), or EUC-JP (Japan). In each of these situations, it is important to back up the UNIX client using a backup application with a language consistent with the selected codepage. Mixing languages (example: having a Japanese UNIX server and a Chinese backup application) may lead to file and directory names that are unreadable. If you do not have a language-consistent backup application, do not back up in UNIX.

6.3 Backing up Unicode-enabled MacOS clients

MacOS 10.1.4 and later support UTF-8 codepage using AFP v3 and later. It is important to back up the MacOSX client with a backup application that supports MacOSx over AFP v3 or later.

7. Other Third-Party Applications

For any application that will be accessing files saved on a Unicode-enabled Snap Server, check with your vendor to make sure you understand the extent of support for the UTF-8 Unicode codepage. In many cases, it may simply be a misrepresented User Interface (UI) display of file and directory names. The application itself may operate correctly.

7.1 Applications bundled with GuardianOS

Snap Servers powered by GuardianOS have bundled applications available for integration into an enterprise. It is important to understand the implications of utilizing these applications once Unicode is enabled on a Snap Server.

7.1.1 NetVault BakBone

The BakBone NetVault v7.1.1 backup application software included in GuardianOS-based Snap Servers is not Unicode-aware; however it is able to display, back up, and restore Unicode named files or directories on operating systems where NetVault supports a Unicode file system.

Currently, a number of limitations exist which are important to note. The following lists the features and functionalities that are potentially impacted by these limitations.

GuardianOS Snap Server file system support

This GuardianOS plug-in is only able to display ASCII characters in both the backup and restore windows. Any other characters will display incorrectly on the GUI. This is a display issue only and does not compromise the backup; all files will be

processed during a backup including the extra attributes attributed to each file.

Include/exclude files

The contents of include/exclude files must only contain ACSII characters. As a result NetVault is not able to include or exclude files or directories that contain non ACSII characters in their names.

Restore renames/relocates feature

When specifying a restore with the “rename” or “relocate” option, only ASCII characters can be used. It is not possible to rename or relocate directories or files to a path or name containing non ASCII characters.

Search in restore screen

The restore screen search feature will not be able to search for non ACSII characters. If these characters are input, the search will not return any result.

Job progress screens

Job progress (monitor) screens will only display ASCII characters.

NetVault logs

Filenames containing non ASCII characters will not be displayed correctly in the logs.

Sorting

Sorting of non-ACSII Unicode entries will not be correct.

Names entered in the NetVault administrative UI

Generally, use of Unicode names for objects created using any UI input screen is restricted. This includes jobs, policies, media labels, and group names. There are some instances where multi-byte characters will appear correctly in the UI, but in such cases the specific platform and codepage for that OS will be the only valid combination for properly displaying file and directory names. The best practice is to always use ASCII characters within the NetVault UI to avoid misrepresented characters.

Interoperability issues

When doing any NetVault domain management on a Snap Server, only ACSII characters will be displayed correctly in the GUI.

7.1.2 NDMP

For environments utilizing the NDMP backup methodology, back ups and restores will work without any file or directory corruption. The Data Management Application (DMA) or backup management application for the NDMP environment will determine Unicode compatibility.

Specific details about the NDMP server that resides on the Snap Server and the NDMP client (DMA) are listed below.

NDMP server (Snap Server-resident software)

The NDMP Server will work with the Unicode file system

correctly during both NetVault back up and restore. The server is not aware the filenames are Unicode, so does not set an attribute for backups indicating the history is in Unicode. This could cause compatibility issues with backup vendors other than BakBone.

NDMP client (DMA backup software)

When defining an NDMP backup, only ACSII paths and file/directory names can be used. The backup index will be created correctly, however only ACSII characters will be displayed correctly in the restore window. Only ACSII paths and file/directory names can be restored. Note: If non-ACSII file/directory names are contained within the path specified for restore, these will be restored correctly; only the file and directory names explicitly selected in the UI used must be ASCII. Rename instructions can only contain ACSII paths and filenames.

For more details using BakBone NetVault, see the section above for information on what to expect with the management of back ups and restores.

7.1.3 Computer Associates eTrust Antivirus

Computer Associates eTrust Antivirus v7.1.3399 is integrated into GuardianOS v4.0.228. At this time the eTrust Antivirus application GUI is not Unicode-aware and does not support any extended characters. While the eTrust Antivirus UI displays unreadable characters for extended characters when Unicode has been enabled, it can still scan files, find viruses, clean viruses, move, and rename virus-infected files. The following is a list of potential issues to be aware of:

- The GUI will not properly display directory or filenames that contain extended characters
- The GUI will not properly display scheduled job names that contain extended characters
- Files that contain extended characters may not be scanned for viruses

File data will **NOT** be corrupted for those filenames that are not displayed properly.

7.1.4 Snap Enterprise Data Replicator (Snap EDR)

When transferring files to and from Windows systems, Snap EDR converts all file and directory names correctly.

When transferring files between UNIX systems, Snap EDR does no character encoding conversion. As such, any set of hosts that are using different codepages will not be able to successfully transfer files without file and directory names becoming unreadable.

Any set of all UNIX systems that use the same codepage can transfer files without issue.

However, in a mixed environment that includes Windows and UNIX systems, where the UNIX system uses a non UTF-8 encodings or more than one encoding, filenames and directory names may appear differently after a transfer. In general,

replications work correctly for ISO 8859-1 characters, but characters outside of the ISO 8859-1 range may not appear properly on the target system (depending on compatibility between the source or target UNIX code page and UTF-8).

Snap Server UTF-8 to Snap Server UTF-8 replications are fully tested and qualified by Adaptec.

7.1.5 Server-to-Server Replication (S2Sv2)

The S2Sv2 replication application fully supports “like-to-like” Unicode codepage replications. For example, if replication data using the UTF-8 Unicode codepage on both the source and the target, the replication will be successful and there will not be a readability issue with the file or directory names. The same holds true if replicating data between two systems that are both using the CP1252 codepage.

The S2Sv2 application translates the CP1252 or UTF-8 GOS file/directory names into Unicode when loaded by JVM. On the wire, the directory and filenames are transmitted in Unicode. At the destination, the directory and filenames are converted back into the codepage used on the destination machine. If the Source machine is configured for Unicode (UTF-8) and the destination machine is configured for non-Unicode (CP1252), any file with non-ASCII characters may be unreadable.

8. Unicode and Expansion Arrays

When an expansion array is converted to Unicode, it stays converted to Unicode. This means that a Unicode-enabled expansion array is only compatible with Snap Server head units that have also been converted to Unicode.

Once an expansion array has been converted to Unicode, it cannot be used with non-Unicode-enabled Snap Servers.

If a non-Unicode expansion array is incorporated into a Unicode-enabled Snap Server, the expansion array will automatically be converted to Unicode.

Caution: Once any Snap Server or expansion array has been converted to UTF-8, there is no way to “un-convert” the file system or disable Unicode without destroying all of your data.

The following example describes expansion arrays and how they operate with Unicode-enabled servers.

Suppose you have a Snap Server 18000 and a Snap Disk 30SA. You enable Unicode on the Snap Server 18000. Both the Snap Server 18000 and the Snap Disk 30SA will be converted to Unicode.

If you later attach the Snap Disk 30SA that was converted to Unicode to a new non-Unicode enabled Snap Server, and then convert the Snap Server to Unicode, the Snap Disk 30SA will be unchanged, as it has been already converted to Unicode. In fact, the expansion array will NOT be incorporated into the new Snap Server until Unicode is enabled on the Snap Server.

9. Summary and Other Issues to Consider

There are some key things to note and remember prior to enabling Unicode on a GuardianOS powered Snap Server.

Many applications that are not Unicode-aware will not be able to correctly display files and folders with extended characters, but may still operate correctly. It’s important to find out the level of Unicode (UTF-8) support from each application vendor prior to enabling Unicode on your Snap Server.

As mentioned before, once Unicode is enabled, there is no way to disable Unicode without destroying all of your data. In the event that Unicode was enabled on a Snap Server and you want to disable it, there is only one way to turn this feature off. It requires that the Snap Server be restored to its factory default state via a maintenance mode operation that will completely destroy all data on the system. If you must disable Unicode you will first need to have a backup of your data prior to enabling Unicode or copy all data off of the Snap Server. You will then need to contact Adaptec Technical Support for instructions on how to execute the maintenance mode operation that will put the Snap Server back to a default state with Unicode disabled.

Note: A backup must be pre-Unicode, as the file system conversion of a Unicode-enabled backup will not translate file and directory names correctly to the non-Unicode-enabled Snap Server. Copying the data off of a Snap Server that will be put back into non-Unicode mode requires that the file and directory names on the Snap Server have not been modified to include any UTF-8 extended characters. If any UTF-8 extended characters exist and are restored back to the Snap Server, those file and directory names will not be readable.

Once Unicode is enabled, there will be a slight performance impact as the file system is now managing the extra translation of the UTF-8 codepage.

Be sure to carefully plan and assess your storage needs for data that will be stored on your Snap Server to determine if enabling Unicode is right for your environment.

adaptec

Adaptec, Inc.
691 South Milpitas Boulevard
Milpitas, California 95035
Tel: (408) 945-8600
Fax: (408) 262-2533

Literature Requests:
US and Canada: 1 (800) 442-7274 or (408) 957-7274
World Wide Web: <http://www.adaptec.com>
Pre-Sales Support: US and Canada: 1 (800) 442-7274 or (408) 957-7274
Pre-Sales Support: Europe: Tel: (44) 1276-854528 or Fax: (44) 1276-854505

Copyright 2005 Adaptec, Inc. All rights reserved. Adaptec, the Adaptec logo, Snap Appliance, the Snap Appliance logo, Snap Server, Snap Disk, GuardianOS, SnapOS, and Storage Manager are trademarks of Adaptec, Inc., which may be registered in some jurisdictions. Microsoft and Windows are registered trademarks of Microsoft Corporation, used under license. All other trademarks used are owned by their respective owners.

Information supplied by Adaptec, Inc., is believed to be accurate and reliable at the time of printing, but Adaptec, Inc., assumes no responsibility for any errors that may appear in this document. Adaptec, Inc., reserves the right, without notice, to make changes in product design or specifications. Information is subject to change without notice.