iSCSI Boot on SPARC Functional Specification

(FWARC 2008/466)

1 Version History

1.0 Revised 11-June-2009

1.1 Revised 12-June-2009

1.2 Revised 21-June-2009

2 Project Description

This provides the OpenBoot code necessary to allow booting Solaris from an iSCSI target.

The code is provided as a package included in the existing network boot package (/packages/obp-tftp), thus offering the iSCSI boot capability to any network adapter which currently provides ordinary tftp boot capability. This slightly differs from the x64 implementation, in that the network card on SPARC is not itself required to contain iSCSI specifics – other than the network driver, all code resides in the platform OBP.

2.1 Architecture Summary

The iSCSI package establishes a TCP/IP connection with the remote target, logs in and authenticates itself, and then provides a conduit for SCSI commands and responses. The existing OBP generic disk driver is used to provide a file system interface. The TCP/IP stack is re-used from the previous project of Wanboot (FWARC 2004/461).

Due to the different semantics of Network devices and Disk devices, the boot command will specify a network device with a series of options specifying the remote target, which will be passed internally to */iscsi-disk*, a device which exports disk device semantics. Secondary booters and Solaris will see a */chosen:bootpath* property specifying */iscsi-disk*, and use that to boot. Solaris must look at other */chosen* properties to determine the underlying location of the root partition just booted.

The target may also be specified in a DHCP *Root Path* option, see RFC 4173. This simplifies the operator's task in only needing to specify “boot net:dhcp” rather than typing the rather long set of details to identify an iSCSI disk. The specifics are documented in §4 (DHCP Interface) below.

3 Bindings

3.1 obp-tftp package modifications, affecting network nodes

The package *obp-tftp* is modified to add support to iSCSI boot, in addition to the current tftpboot and wanboot. This primarily involves providing information parsed from device arguments and opening an

interface to tcp/ip for iSCSI's use. The interface to tcp/ip is private to iSCSI, not intended to be used by other applications.

3.1.1 obp-tftp Methods

The following new methods are specific to iSCSI support, and are not normally even visible to users. These are documented here simply because they are presented as *external* from *obp-tftp*. These methods all presume only one connection exists per device instance so there is no need for references indicating which connection is being addressed.

3.1.1.1 iscsi-connect (--)

Establishes the TCP connection with the remote target. The information needed to determine where to connect has been provided on the network-device arguments or obtained from DHCP.

3.1.1.2 iscsi-read (addr len -- actual)

Performs a TCP read from the remote target over the connection opened by *iscsi_connect*.

3.1.1.3 iscsi-write (addr len -- actual)

Performs a TCP write to the remote target over the connection opened by *iscsi_connect.*

3.1.1.4 iscsi-disconnect (--)

Breaks the TCP connection and reclaims resources allocated.

3.1.2 obp-tftp Device Arguments

iSCSI support in the obp-tftp package supports a series of device argument keywords to identify the destination iSCSI target, following the “keyword=value” format described in FWARC 2002/461 and FWARC 2004/579.

A network device path with these keywords may result in a tremendously long boot command. Note that the command line (including expanded “*net*” devalias) may not exceed 256 characters. Should arguments exceed that length, the arguments must be placed in the *network-boot-arguments* NVRAM variable.

All of the below arguments (or the values provided by DHCP or internal defaults) will be stored as properties in */chosen*, see §3.2.1 (/chosen Properties) below.

3.1.2.1 iscsi-target-ip

Dotted-decimal formatted IP addresss (e.g., 255.255.255.255) of remote iSCSI target.

3.1.2.2 iscsi-target-name

iSCSI qualified name of desired disk target, as described in RFC 3720 §3.2.6.3 (iSCSI Name

Structure). This will usually be a string in the form “iqn.1986-03.com.sun:02:...”.

3.1.2.3 iscsi-port

Optional decimal formatted integer from 1 to 65535, representing the TCP port number of the remote iSCSI implementation. This will default to 3260 if not specified.

3.1.2.4 iscsi-partition

Optional string specifying bootable partition on the target. Defaults to null string if not specified. The secondary booter will later take a null string to mean partition “a”.

3.1.2.5 iscsi-lun

Optional hexadecimal dash-separated field encoding the 64-bit LUN of the remote target. Will default to zero if not specified.

In its trivial manifestation (lun number smaller than 65535), this is simply a hexadecimal encoding. For larger values, the formatting of this 64-bit field is complex. We quote in its entirety the paragraph from RFC 4173 §5 describing how this field must be formatted:

The "LUN" field is a hexadecimal representation of the LU number. If the LUN field is blank, then LUN 0 is assumed. If the LUN field is not blank, the representation MUST be divided into four groups of four hexadecimal digits, separated by "-". Digits above 9 may be either lower or upper case. An example of such a representation would be 4752-3A4F-6b7e-2F99. For the sake of brevity, at most three leading zero ("0") digits MAY be omitted in any group of hexadecimal digits. Thus, the "LUN" representation 6734-9-156f-127 is equivalent to 6734-0009-156f-0127. Furthermore, trailing groups containing only the "0" digit MAY be omitted along with the preceding "-". So, the "LUN" representation 4186-9 is equivalent to 4186-0009-0000-0000. Other concise representations of the LUN field MUST NOT be used.

3.1.2.6 iscsi-initiator-id

Optional string containing iSCSI name of initiator. This is a string similar to the *iscsi-target-name*, this time specifying the initiator (boot prom) side of the connection. If not specified, a host-unique default will be provided. The default assumed will be of the form:

`iqn.1986-03.com.sun:boot.<system-mac-address>`

3.2 /chosen node

3.2.1 /chosen Properties

These properties reflect either device arguments specified to the boot command, values obtained through DHCP or RARP, or defaults. They are always created (along with *host-ip, router-ip* and

subnet-mask described in FWARC 2002/561) during an iSCSI boot.

3.2.1.1 iscsi-target-ip

Type: Prop-encoded array, property-encoded string.

Contents: Dotted-decimal formatted IP addresss (255.255.255.255) of remote iSCSI target.

3.2.1.2 iscsi-target-name

Type: Prop-encoded array, property-encoded string.

Contents: iSCSI qualified name of desired disk target, as described in RFC 3720 §3.2.6.3 (iSCSI Name Structure).

3.2.1.3 iscsi-network-bootpath

Type: Prop-encoded array, property-encoded string.

Contents: The complete *device path* to which the *device-specifier* of the iSCSI boot command was resolved.

3.2.1.4 iscsi-port

Type: Prop-encoded array, property-encoded string.

Contents: Decimal formatted integer from 1 to 65535, representing the TCP port number of the remote iSCSI implementation. This will default to 3260 if not specified.

3.2.1.5 iscsi-partition

Type: Prop-encoded array, property-encoded string.

Contents: Optional string specifying the bootable partition on the target.

3.2.1.6 iscsi-lun

Type: Prop-encoded array, property-encoded string.

Contents: hexadecimal dash-separated field encoding the 64-bit LUN of the remote target. See §3.1.2.5 above for more details on the formatting.

3.2.1.7 iscsi-initiator-id

Type: Prop-encoded array, property-encoded string.

Contents: iSCSI name of initiator. This is a string similar to the *iscsi-target-name*, this time specifying the initiator (boot prom) side of the connection. The format is specified in RFC 3720 §3.2.6.3. If not specified, the initiator id will default to:

`iqn.1986-03.com.sun:boot.<system-mac-address>`

3.3 /iscsi-disk node

The */iscsi-disk* node is a psuedo-node providing an interface to export disk semantics while data is being transported over a network interface. The node itself contains code behaving like a SCSI HBA, exporting methods obeying the binding in IEEE 1275 Annex E, SCSI Host Adapter package. On *open*, this code interposes */packages/idisk* - see §3.4 (/packages/idisk) below. This interpose results in presenting a scsi *disk* interface without requiring an additional device tree node,

When a system is booted using iSCSI, the boot command must be presented with a network node with arguments, but the secondary booters and host operating system will see only that the */chosen*:*bootpath* property refers to something like "*/iscsi-disk:a"* (the argument here is obtained from the *iscsi-partition* keyword).

3.3.1 /iscsi-disk properties

3.3.1.1 obp-tftp-ihandle

Type: Prop-encoded array, encoded with *encode-int*.

Contents: *ihandle* of the *obp-tftp* package, at the time of open.

This is a temporary property, to convey hidden information from *obp-tftp* to */iscsi-disk.* The property is created during the open of the network device and is deleted at the completion of the */iscsi-disk* open.

3.4 /packages/idisk

A new package */packages/idisk* is added. This package is essentially the existing scsidisk.fth, but packaged in a form which can be interposed (see IEEE 1275 working group proposal #272). This permits a single node which has the SCSI HBA semantics defined to interpose the *idisk* package and present a standard disk interface without requiring an additional device tree node.

3.5 Security Keys

3.5.1 User Interface

Part of the iSCSI protocol includes CHAP authentication, to ensure we are reaching the correct disks (see RFC 3723, §2.4.1). This authentication protocol contains a username and password, which must be kept private to prevent spoofing. During the boot process, these are obtained from Key Storage (see FWARC 2002/182), which is a form of NVRAM variable not exposed to general users. Solaris and secondary booters may obtain the keys with the OBP Client Interface *SUNW,get-security-key*, and thus use the same authentication parameters as the initial boot.

3.5.1.1 set-ascii-security-key (“key-name< >key-value<eol>” --)

key-name contains the name of the key (see security key registry), up to 64 characters long.

key-value contains the ASCII characters corresponding to desired key value, up to 64 characters long.

This is a parallel interface to the *set-security-key* method defined in FWARC 2002/182 and 2003/143. The original method allows only hexadecimal input, this method accepts ASCII strings. Since iSCSI passwords are usually specified in ASCII, hexadecimal input is both overkill and inconvenient.

This interface does not provide the ability to enter all possible key values – the use is restricted to keys composed of standard ASCII characters allowed for Forth word definitions. For keys containing more exotic characters, the string must be converted to hexadecimal and the original *set-security-key* method must be used.

3.5.2 Security Key Registry

The following two Security Key names are added to the Security Key registry.

3.5.2.1 chap-user

This is the username (sometimes known as short name, userid or simply name). which is transmitted over the network in the clear and is thus not considered strictly secret. In Solaris' *iscsiadm* this is set with the *-H* or *--CHAP-name* switch. It may be anywhere from 1 to 16 characters in length. We are maintaining this name as a security key simply to keep it together with the below *chap-password*, simplifying code by providing a single method used to retrieve both values. This will usually be set with a command such as:

```
ok set-ascii-security-key chap-user admin
```

3.5.2.2 chap-password

This is the authentication password, never transmitted over the network, which must be from 12 to 16 characters long. This is sometimes known as the CHAP secret or CHAP password. In Solaris' *iscsiadm* this is set with the *-C* or *--CHAP-secret* switch. This value is used as the hidden value in exchanging of public encrypted values to validate that both sides are using the same password. This will usually be set with a command such as:

```
ok set-ascii-security-key chap-password abcdef0123456789
```

4 DHCP Interface

A DHCP server may be configured to provide the iSCSI boot information required, with a *Root Path* option. This is specified in detail in RFC 4173, the summary of the format is:

```
iscsi:<servername>:<protocol>:<port>:<LUN>:<targetname>
```

The above string must start with the literal six characters “iscsi:”, followed by the five variables separated by colons.

The <*servername*> will be specified as an IP address (since OBP doesn't know about name-address

translations) in dotted decimal form, as all openboot IP addresses are specified.

The <*protocol*> is required to be "6", indicating TCP.

The <*port*> is a decimal string indicating the IP port used to communicate with the iSCSI server – usually 3260.

The <*LUN*> indicates the SCSI LUN holding the boot disk, in dashed hexadecimal format, as specified in §3.1.2.5 above.

The <*targetname*> is the long string used to identify iSCSI targets, as specified in RFC 3720 §3.2.6.3, ISCSI Name Structure (See also RFC 3722). An item to note is that this *Root Path* option consists of colon-separated items, and the last item contains at least one colon (and in some cases it will contain two). This works only because it is the last item in the DHCP variable list.

5 *show-iscsi* command

The *show-iscsi* command is provided as a debug facility, to inquire from the iSCSI target server which target-ids are available, if the server provides SendTargets capability (RFC 3720 §12.3).

For convenience, the *show-iscsi* command takes the same format argument as a manual configuration boot command. It ignores the *iscsi-target-id* keyword (which is nonetheless required to indicate an iscsi operation), and prints a list of valid target-ids and corresponding LUNs on the specified target. This format is convenient for the case where a non-DHCP boot command fails, the exact same argument(s) may be provided to the *show-iscsi* command to determine what the proper target name on the destination IP address could be.

6 Source Components

There are three major source directories affected by this work:

- obp/pkg/netinet – the existing tftp and http boot (wanboot) package, which exists in the running OBP under */packages/obp-tftp,* receives minor modifications to support iSCSI.
- obp/pkg/iscsi – a new module, the iSCSI code itself, which is placed in the running OBP under */iscsi-disk.*
- obp/pkg/idisk – a new module, interposable disk, which contains essentially no code beyond the existing obp/dev/scsi/targets/scsidisk.com. It is placed in the running obp under */packages/idisk*.

7 Description of iSCSI boot internal mechanics

The semantics of network boot (where *open* can perform no I/O, only *load* may initiate activity to the device) compared to disk boot (where *open* must read disk labels and determine partitions before completing) have required us to use two items in the device tree to carry out a boot from an iSCSI disk – a network device and the psuedo-device */iscsi-disk*.

The user must specify a boot string based on the network device which connects to the iSCSI target; either "boot net:dhcp", or the manually specified keyword sequence "boot net:iscsi-target-id=...". The network driver brings in the *obp-tftp* package, which parses the keywords. When the *load* method is

invoked by the *boot* command, *obp-tftp* will proceed to ask RARP or DHCP for more information if specified, and then proceed to carry out one of three forms of boot: tftp, http or iSCSI.

In the case of iSCSI, after creating the */chosen* properties specified in 3.2.1 above, it will open the */iscsi-disk* psuedo-node to start the disk boot process. The driver for the */iscsi-disk* psuedo-node will obtain the information about the iscsi target from properties in */chosen* and initiate the TCP/IP connection to the iSCSI server. Once a connection has been established, the */chosen:bootpath* property will be re-written specifying */iscsi-disk* as the boot device – this causes Solaris to find a disk device, and produce file I/O requests to that disk device rather than trying to perform network activity over a network device.

The secondary booter *ufsboot* is free to re-write the arguments to */iscsi-disk*, which it does as part of its file-system initialization (usually an empty argument list is replaced with ":a"). Solaris will see a boot path consisting of "/iscsi-disk:a" as a result.

8 References:

- RFC 3720 – Internet Small Computer Systems Interface (iSCSI)
- RFC 3721 – iSCSI Naming and Discovery
- RFC 3722 – String Profile for iSCSI names
- RFC 3723 – Securing Block Storage protocols over IP
- RFC 3783 – Command Ordering considerations with iSCSI
- RFC 4173 – Bootstrapping clients using iSCSI
- RFC 3980 – T11 NAA Naming format for iSCSI
- RFC 5048 – iSCSI Corrections and Clarifications
- FWARC 2002/182 – Key Storage Signature Verification
- FWARC 2003/143 – Extension of FWARC 2002/182, Security Keystore
- FWARC 2002/561 – OBP-TFTP Wanboot Extensions
- FWARC 2004/579 – Miscellaneous Wanboot Changes
- PSARC 2008/427 – iSCSI support in Solaris