CRITICAL I/O WHITE PAPER

Using StoreEngine & StorePak as a High Speed Data Recorder

Abstract

The StoreEngine Data Recorder provides ultra high speed and scalable data recording capabilities, with a wide variety of data source options. The recorder's basic building blocks consist of the StoreEngine unified storage blade with its onboard SATA SSD storage, and the optional StorePak removable SATA storage modules, all interconnected through a high performance PCIe fabric. Combined with StoreEngine's turnkey Recording Mode software, this provides a complete data recorder that can be implemented with as little as a single blade, and with no custom software required. StoreEngine based recorders are easily scaled in capacity, bandwidth, and channels by adding additional StoreEngine or StorePak blades. The Recording Mode software automatically extends the PCIe data fabric and aggregates storage capacity and performance of the blades.

Revised 5/3/12

StoreEngine Data Recorder

The StoreEngine Data Recorder consists of data recorder hardware and software that provides a flexible and highly scalable recording platform to continuously record high bandwidth data streams from ADCs, FPGAs, video streams, processor boards, and other sources. StoreEngine based recorders can support real time recording of up to 5 GByte/sec or higher, with recording capacities of up to 12TB (Terabytes) or more. The StoreEngine recorder platform provides a simple, scalable recording architecture through aggregation of multiple StoreEngine and StorePak storage blades that maintains a single entity operational and management view.

Core to the recorder is the StoreEngine Recording Mode software, which runs directly on StoreEngine blades providing turn-key data recording operation. The Recording Mode software coordinates the striping of data across multiple StoreEngine and/or StorePak blades. It also coordinates unified playback of recorded data, so the multiple StoreEngines and StorePaks in the data recorder appear as a single data source when recorded data is replayed. All of the functionality to control both the recording of data and the play back of data from multiple blades is built into the recording software. Utilizing Critical I/O's Recording Mode software greatly reduces the effort required to field a scalable wideband data recorder.

StoreEngine and StorePak

StoreEngine and StorePak are flexible storage building blocks that can be used to implement a wide range of data storage systems. **StoreEngine** is an ultra-high performance *Storage Controller* VPX blade (also VXS) that hosts up to 1.5TB of non-removable on-board SSD storage. It simultaneously provides high performance recording functionality, serves block data (like a disk drive or RAID system), as well as provide NAS file sharing (like a NFS/CIFS file server).

StorePak is a *Storage* expansion blade that can host up to 3TB of easily removable and hot swappable SSD storage.

StoreEngine, and optional StorePaks, provide unmatched storage capability, ultra high performance and high capacity all within a small size, weight, and power (SWaP) footprint. StoreEngine is ideal for high bandwidth embedded data recording, NAS file serving, and general purpose RAID applications. Systems are easily scalable in capacity and performance by simply adding additional StoreEngine and/or StorePak blades.

While StoreEngine's other operating modes (NAS file sharing or DAS/RAID) can also be used in data recording applications, this paper focuses on StoreEngine's *Recorder Mode*, which provides the user with the highest performance, most scalable, and simplest data recording solution.

Data Recorder Overview

The StoreEngine Data Recorder mode provides a capability to record multiple streams of data directly from a wide variety of PCIe connected "data sources" at rates of up to 1.5 GB/s per storage blade. Multiple storage blades can be combined for higher aggregate recording rates. Recorded data may be read back to PCIe connected "data destinations", or alternatively recorded data may be played back via Fibre Channel, or via 1/10 Gb Ethernet connections from the StoreEngine using standard NFS, FTP, or CIFS file access.

All Data Recorder functionality is accessed and controlled through a single control interface to a single StoreEngine board. Most often, this access and control is provided through a Recording Driver that is hosted on a user's System Controller SBC, which is PCIe connected to StoreEngine. Web and network based control interfaces are also available; these are most often utilized when recording from directly from ADC or Ethernet data sources.

The StoreEngine and StorePak blades, along with the data source(s), are typically hosted in a VPX rack, interconnected using a standard mesh or PCIe switched backplane. A VPX mesh backplane provides point-to-point "fat pipe" PCIe connections between boards, while a switched backplane uses a PCIe switch board to provide slot to slot connectivity.

StoreEngine and StorePak provide a highly configurable and partitionable PCIe switched front-end that allows them to be customized to fit a wide variety of system architectures. The VPX version of StoreEngine in particular supplies rich PCIe connectivity -- up to seven PCIe Gen2 x4 ports are available for backplane and RTM PCIe access. StorePak provides up to six PCIe Gen2 x4 boards backplane connections per blade.

Recording Data Source Options

Options for the recorded data source range from a data stream from a simple PCIe connected ADC, to a more intelligent PCIe connected device such as an FPGA processor, or a PCIe connected standard CPU board (SBC). The source may also consist of 10GbE network connection to record an Ethernet or UDP/IP data stream.

Simple PCIe Source (example: ADC)

For simple PCIe sources like ADCs, the StoreEngine feeds the PCI source with a sequence of PCIe addresses and block sizes. The PCIe source (ADC) then DMAs blocks of data to the supplied addresses, which actually point to memory buffers hosted on the StoreEngine(s). Simple recording sources like ADCs can be completely controlled by the StoreEngine recorder.

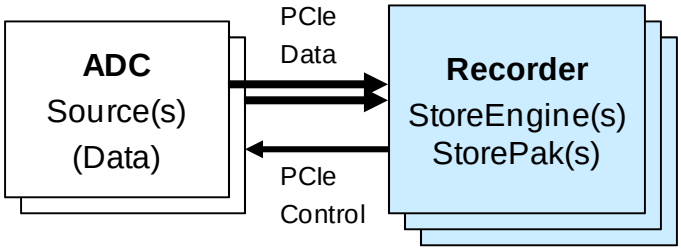

Figure 1. Simple PCIe data source

Intelligent PCIe Source (example: FPGA signal processor board)

Intelligent PCIe sources differ from simple PCIe sources in that an external controller element (typically some sort of processor board, labeled System Controller in the diagram below) is assumed. The system controller function controls the writing/reading of data to/from the recorder via commands that are sent from the system controller to the recorder through the Recorder Mode driver hosted on the system controller board.

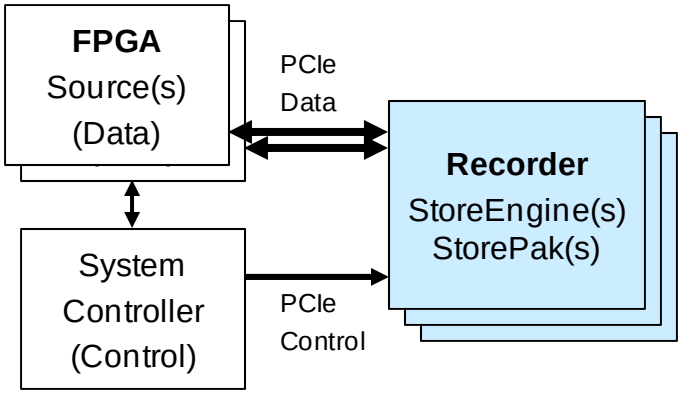

Figure 2. Intelligent PCIe data source

Processor Source (example: PPC or x86 CPU board)

In this model, a user's processor board runs a light weight Recorder Driver, which simply coordinates the DMA transfer of large blocks of data directly from host memory to the recorder. This is a highly efficient method of recording large volumes of data from a processor board that relieves the process from the CPU intensive task of managing low level storage.

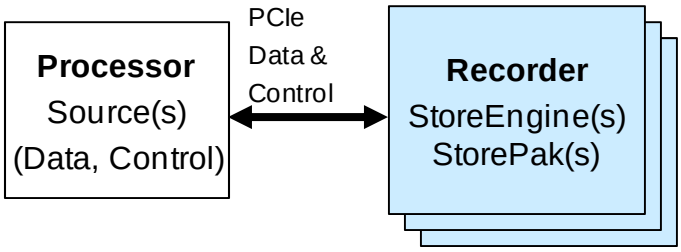

Figure 3. Processor data source

For processor data sources, it is important to highlight the fundamental differences in operation between *Recording Mode and DAS Mode* (which is also supported by StoreEngine).

In *Recording Mode*, the StoreEngine completely manages its storage, and writes the data blocks to disk using RAID 0, using a recording file system. This StoreEngine recording file system supports both buffered mode and direct mode operation, and also coordinates striping data from a single data stream across multiple StoreEngines/StorePaks for increased capacity and performance. Typical performance is about 1 GB/s per blade, with higher levels of performance available by aggregating additional blades.

By contrast, in *DAS mode*, the StoreEngine just aggregates its raw storage, and appears to the host processor board as one (or several) big disk drives. The host board must run a normal PCIe DAS host driver, and must fully control the allocation and use of the raw storage blocks presented by the StoreEngine, often done using a standard file system (e.g. ext3, ext4) or custom user developed software. Because the host must implement some sort of file system to allocate and manage storage, typical DAS based recording performance is lower than Recording Mode, about 300 to 600 MB/s.

Ethernet Source (example: UDP data stream)

The 10Gb Ethernet recording architecture leverages a 10GbE RTM (or XMC) to provide the 10GbE interface(s). The 10GbE interface is completely controlled by the Recording Mode software, which can be configured to record raw Ethernet traffic, or more commonly, a UDP data stream representing sensor or video data.

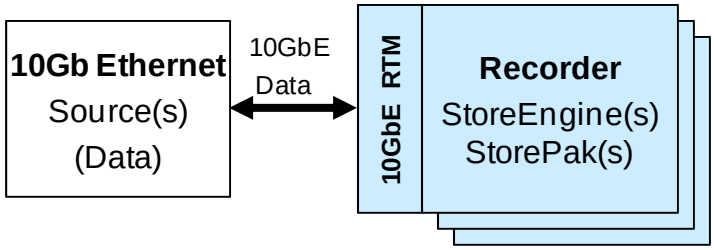

Figure 4. Ethernet data source

Recording Modes – Direct vs. Buffered

Direct mode is a way of using StoreEngine/StorePak that results in data being transferred directly from PCIe data sources to/from the storage resources hosted on StoreEngine or StorePak modules, without first being buffered in StoreEngine memory. This contrasts with Buffered mode, where data is buffered in a StoreEngine prior to being moved to the storage resource. The two different modes are illustrated below in figures 5 and 6.

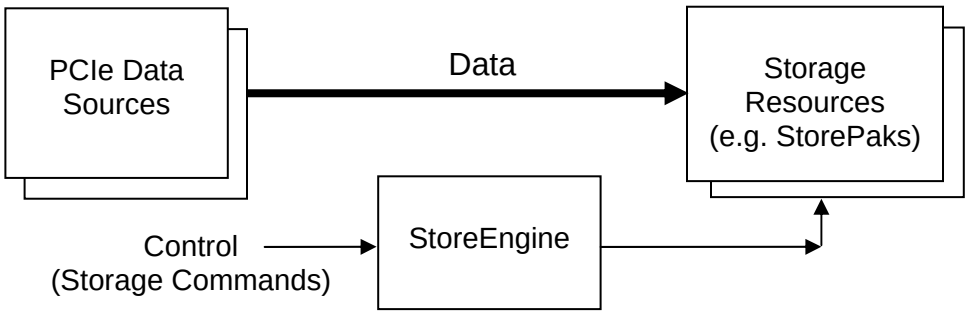

Figure 5. Direct Mode Operation

Direct mode has the advantage of higher performance, because data is not first buffered in a StoreEngine. Direct mode requires that the storage resource (e.g. StorePaks) be able to read data from PCIe accessible memory on the PCIe data sources. This is sometimes referred to as a “pull” data transfer model. Direct mode has a further advantage of increased storage density, as multiple StorePak modules may be controlled using a single StoreEngine module. Only RAID 0 is supported in Direct mode.

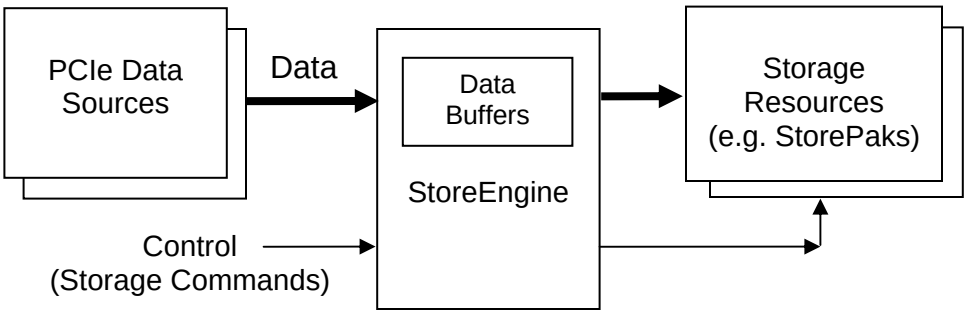

Figure 6. Buffered Mode Operation

In contrast, buffered mode operation requires that data first be buffered in a StoreEngine. Buffered mode has the advantage of flexibility. PCIe data sources may "push" data into data buffers on the StoreEngine(s), rather than relying on StorePaks/StoreEngines to "pull" data. The timing of these data transfers is more flexible, and can be controlled by the data source. StoreEngine transfers data to the Storage Resources on a decoupled timeline. Both RAID 0 and RAID 5 are supported in buffered mode.

The StoreEngine recorder supports concurrent use of direct mode and buffered mode operation. For example, direct mode may be used to record streams of data sourced directly from PCIe data source devices, while buffered mode may be used concurrently to record data streams sourced from a System Controller SBC.

Examples: Simple PCIe (ADC/FPGA) Data Recording (Buffered Mode)

For ADC/FPGA recording from "simple" PCIe devices, the ADC or FPGA function is typically hosted as either 1) a VPX ADC/FPGA module with single or multiple PCIe interfaces, or 2) an ADC/FPGA XMC module hosted on VPX XMC carrier, possibly with an integrated PCIe switch. Two typical PCIe (ADC/FPGA) recording configurations are illustrated in figures 7 and 8. Note that while the figures illustrate specific PCIe connection topologies, many other topologies can also be used.

Figure 7 shows two ADC sources connected to a StoreEngine recorder via a PCIe switch blade, while figure 8 shows an FPGA source connected directly to the StoreEngine recorder via point to point backplane connections.

In this "simple" PCIe model, the data source ADC (or other PCIe source device) must have the ability to perform block DMA, based on a list of addresses and block sizes are provided to the ADC or FPGA PCIe DMA interface. Occasional additional "non-data" transfers to other defined addresses (mailbox interrupt addresses) are used by the StoreEngine blades to determine when a series of block data DMAs has been completed.

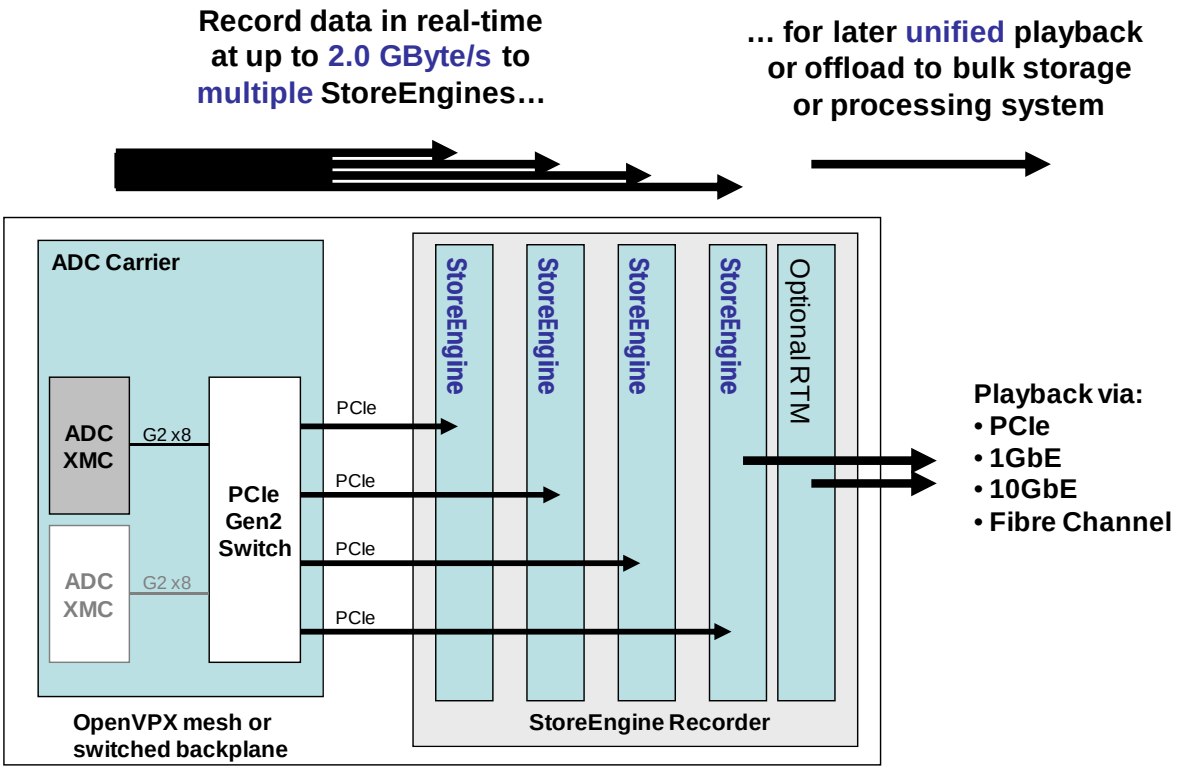

Figure 7. StoreEngine recorder (buffered mode) driven from an ADC data source.

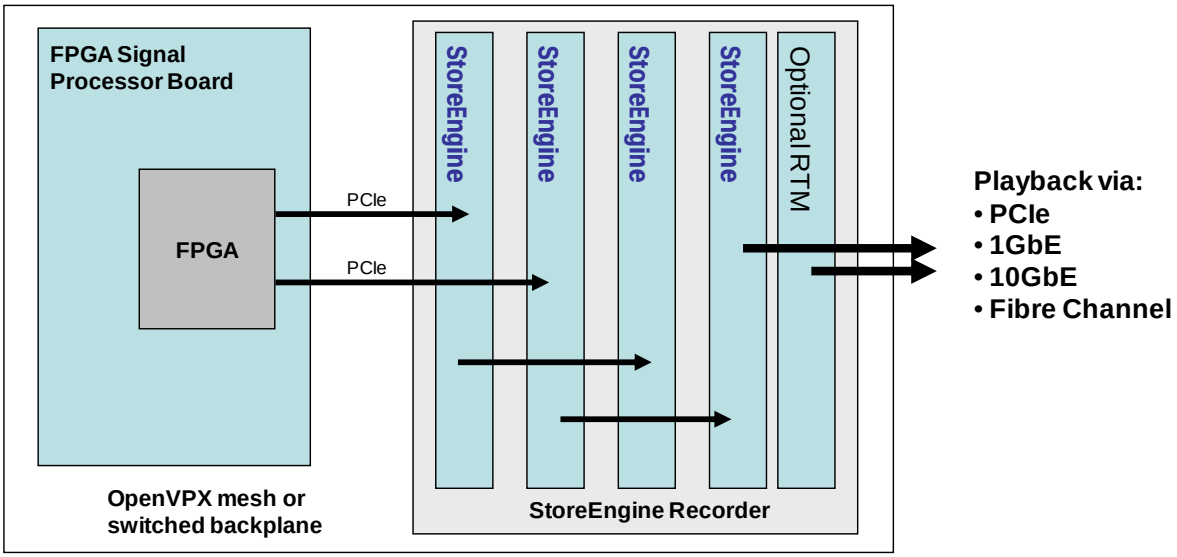

Figure 8. StoreEngine recorder (buffered mode) driven from an FPGA data source.

Example: FPGA Recording Architecture (Direct Mode)

Figure 9 shows a direct mode recorder that uses a combination of StoreEngine blades and StorePak blades. This four channel recorder supports an aggregate recording rate of 5 GByte/s, and a total hot-swappable recording capacity of 12TB. In *direct mode*, on StoreEngine controls four StorePak blades, which "pull" data directly from the FPGA PCIe data sources. This is an example of an "Intelligent" FGPA based PCIe data source, where the FPGAs data source(s) are controlled by the user's System Controller SBC, which also coordinates writing/reading data to/from the StoreEngine recorder.

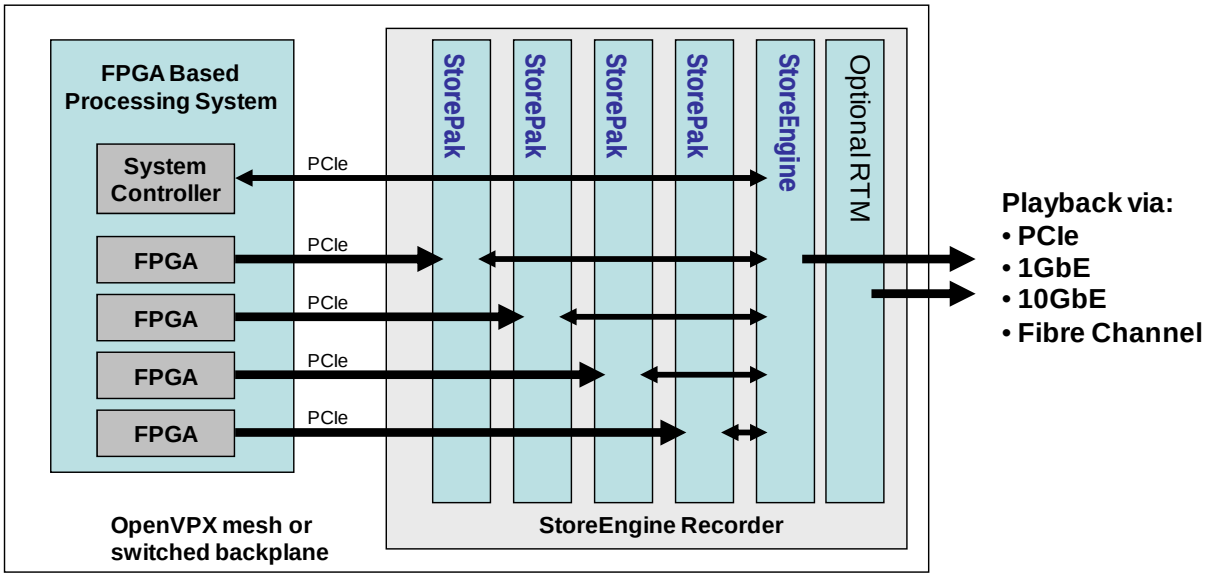

Figure 9. StoreEngine recorder (direct mode) driven from four FPGA data sources.

Example: Ethernet Recording Architecture (Buffered Mode)

For 10 Gb Ethernet recording, a 10 Gb Ethernet NIC function is typically hosted either as:

- A 10GbE NIC Rear Transition Module (RTM)
- A 10GbE NIC XMC module hosted on VPX XMC carrier

A typical Ethernet recording configuration is illustrated in figure 10. The 10GbE NIC is controlled by the Recording Mode software. The software aggregates buffer addresses from all of the StoreEngines in the recorder, and supplies these buffer addresses to the 10GbE NIC.

The NIC then pushes data segments to the StoreEngines, based on the DMA address lists produced by the StoreEngine. Each data segment consists of either a raw Ethernet frame, or one of a sequence of UDP datagrams that comprise a UDP data stream. The StoreEngine blades group these segments into larger slices, and then use RAID 0 to distribute the slices of data among their local SATA drives.

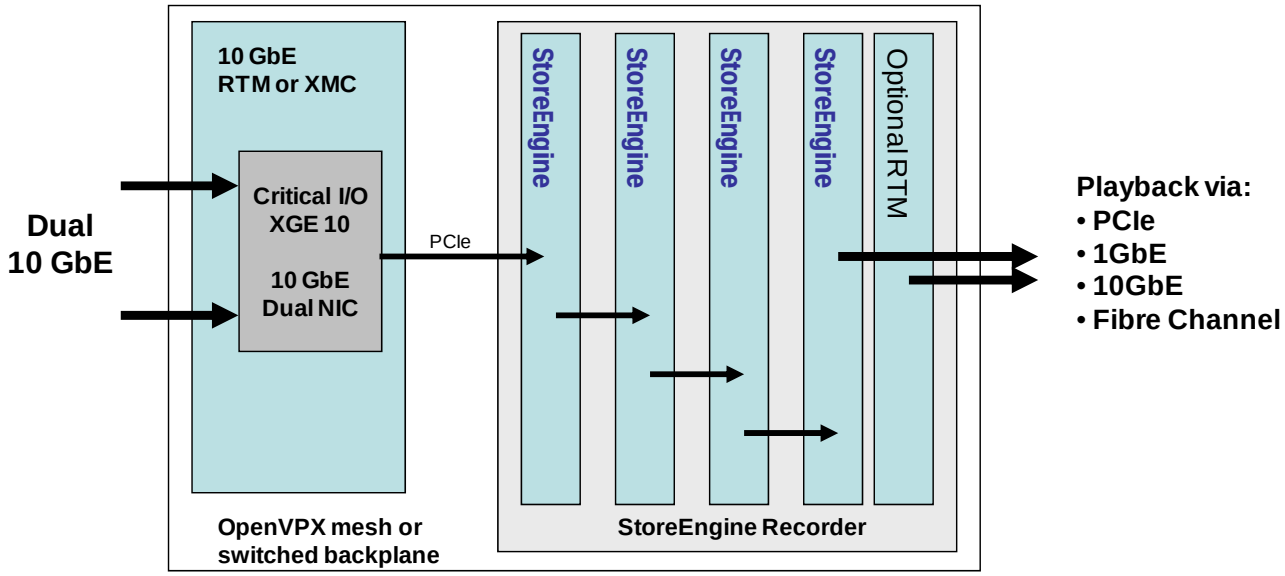

Figure 10. StoreEngine 10GbE recording architecture.

Recorder Data Management and Metadata

The Recording Mode software stores recorded data on the blades SATA storage drives as a series of "files". A file is defined as a continuous recording of data between a "start" command and a "stop" command (equivalent to an "open" and "close"). Each file may be distributed among all of the StoreEngine blades that comprise the data recorder, and are further distributed across all of the SATA drives within each StoreEngine blade. Files are recorded contiguously and can only be deleted en masse. However, recorded data can be replayed or extracted on a file by file basis.

A portion of each SATA drive is also reserved for metadata. System metadata includes information about the recorder configuration, such as the number of blades, and the number and size of the SATA drives on each blade. File level metadata includes the recording size, the start and stop time, and the starting location on disk, as well as precise time-stamping of each recorded block. File metadata may be extended to include additional information such as GPS information or other application specific metadata.

Recording Playback / File Extraction

Several methods are available to access recorded data. Intelligent PCIe connected devices such as FPGAs or processor boards may simply read the previously recorded files by issuing read commands through the recording mode driver. This method provides read performance that is equal to or better than recording performance, and is the simplest "playback" method to use in these types of systems..

Recorded data may also be "played back" via a normal PCIe DAS host connection; via a Fibre Channel DAS connection; or via a 1/10 Gb Ethernet connection. In this playback model, the device receiving the playback data is termed a Playback Consumer. The Playback Consumer may be a processor board, an external PC, or any other computer that can connect to the data recorder via PCIe, Ethernet, or Fibre Channel.

To any Playback Consumer, the multiple StoreEngines and StorePaks that comprise the recorder appear as a single data source. The recorder coordinates responses to Playback Consumer requests by pulling the necessary slices of data from each storage blade, and reassembling the slices into a single stream that is provided to the playback consumer. The interface allows the selection of the file to be played back, and also allows for the specification of playback range.

For 1 or 10 Gb Ethernet playback, the recorder creates and exports a pseudo file system, where each file in this file system is mapped to an aggregated recorded file. This allows the use of standard network file sharing protocols such as FTP or NFS to access the aggregated recorded data. The recorder also supports a unique UDP based ultra-high performance 10GbE file access protocol that allows recorder files to be extracted at sustained rates of 1GB/s using a standard 10GbE interface.

Data Recorder Configuration and Management

Three methods of configuring and controlling StoreEngine recorder are available. All methods allow users to configure and monitor the recorder, and start or stop recording sessions. For all methods, a single control interface to one StoreEngine controls all of the StoreEngines and StorePaks which comprise the recorder.

For initial recorder system configuration, the built-in web based management function (that resides on StoreEngine) can be used to configure the Data Recorder, and can subsequently be used to monitor the StoreEngine and StorePak boards, and associated recorder functionality

Recorder Driver control interface.

This method is used with intelligent PCIe data sources such as ADCs, or with Processor data sources. In both of these cases a user supplied processor board hosts a Critical I/O supplied Recording Mode driver. The Recorder Driver interface allows the recorder to be initialized and configured, and allows recording sessions to be started and stopped with a variety of recording options. The recording mode driver supports multiple streams of data, and multiple channels within each stream. The writing and reading of blocks of data to/from each channel and stream may be individually controlled through the recording mode driver.

Web based recording control interface.

This method is most useful with “simple” PCIe data sources such as ADCs, or with Ethernet data sources. The web management interface allows the recorder to be initialized and configured, and allows recording sessions to be started and stopped with a variety of recording options.

Network message based recording control interface.

The network based control method allows any computer or processor board on the same Ethernet network to send control messages to the recorder using a simple UDP socket based protocol. All configuration, control, and status functionality is available using this method.

Recorder Configuration Functions

The following user selectable recording configuration functions are available.

- CONFIG - The button will format the data recorder, sets the block size and places the data recorder in a state where it can start recording data

- RESET -This button initiates a purge of all recorded data and reinitializes the recorder with the current settings

Recording Control Options

When recording from a simple PCIe (e.g. ADC) or Ethernet (e.g. UDP stream) source, the following user selectable recording control options are:

- Limit Recording Time – This option allows the user to specify a specific recording time limit. Once the recording is started, it will be automatically stopped at the specified time limit.

- Limit Recording Size - This option allows the user to specify a specific recording file size. Once the recording is started, it will be automatically stopped once the recorded data set reaches the specified size.

- Block Size – This option specifies the block size that is used for data transfers from PCIe recording sources (ADC/FPGA/Processor).

- Recording Source Selection – This option selects the source of the recording data stream. Options include PCIe (ADC/FPGA), raw Ethernet, or UDP streaming Ethernet.

- Continuous Overwrite Mode – This option allows data to be recorded continuously. Old data will be overwritten with new data when the storage resources are full.

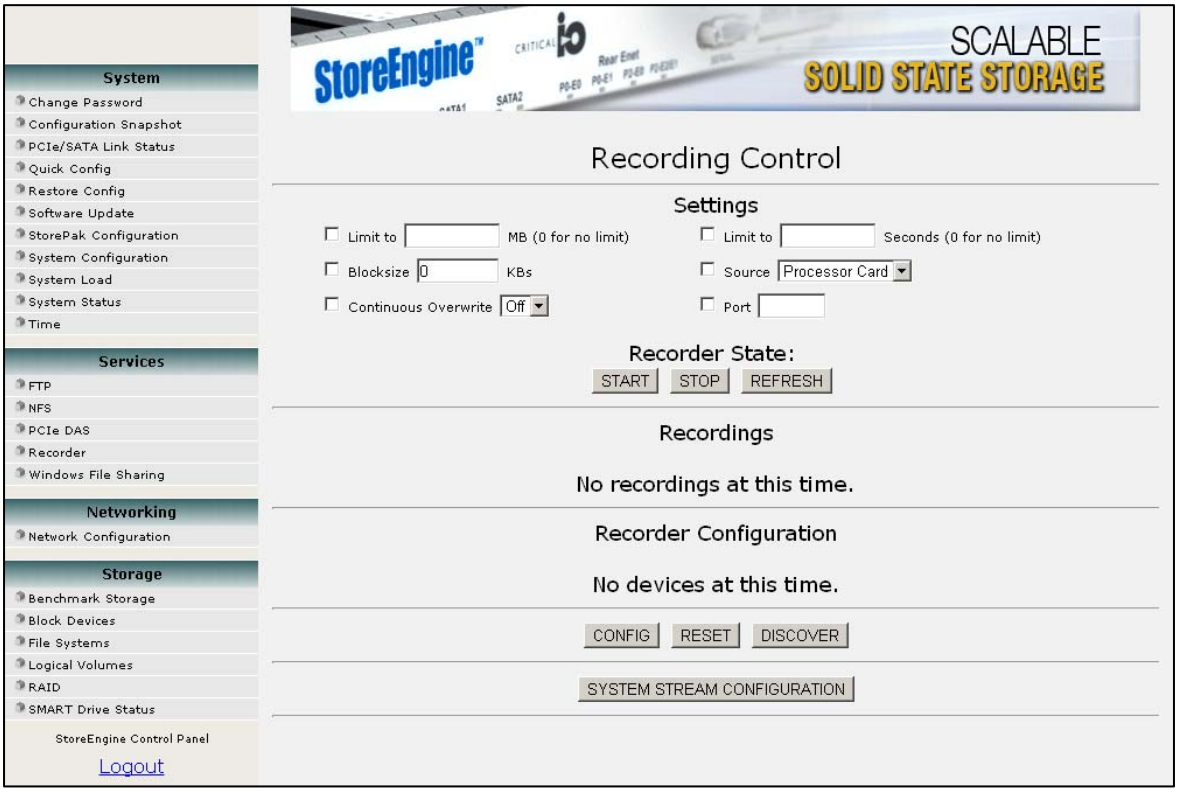

Figure 11. StoreEngine recorder web control page for simple PCIe sources.

Figure 11 shows an example of the recorder configuration web management page as used when recording from a "simple" PCIe source such as an ADC. In this example, the recorder configuration is small; only a single StoreEngine. All of the same configuration, control, and status capabilities can also be accessed either via the web management interface, or through use of network messaging management interface.

Storage Media Choices

The performance, capacity, and reliability characteristics of the recorder are strongly influenced by the choice of storage media. Key media characteristics are defined by five main parameters:

- Capacity – Raw storage capacity of a drive, measured in GB orTB.
- Performance – Sequential and/or random read/write performance of the media, measured in MByte/s and/or IOP/s.
- Write endurance – How many times the media can be fully overwritten before a significant number of storage regions become unusable or performance degrades.
- Reliability – The probability of a drive or media failure such that some portion of data becomes inaccessible.

The types of SSD storage commonly used for storage blades are:

- SSD-SLC – Solid State Drive, Single Level Cell (SLC) Flash Media
- SSD-MLC – Solid State Drive, Multi-Level Cell (MLC) Flash Media

For recording applications, the media choice depends on the particular application. For high bandwidth continuous recording, where the media is continuously overwritten with new data, SSD-SLC media may be the best choice. For high bandwidth high capacity applications, SSD-MLC media is generally the best choice.

Table 1. Storage media characteristics

Media Type	Sequential Performance	Random Performance	Write Endurance	Relative Reliability	Relative Capacity	Relative Cost
SSD-SLC	Excellent	Excellent	High	High	Lower	Highest
SSD-MLC	Excellent	Moderate	Lower	Lower	High	Moderate

Table 2. Fixed and removable storage capacities

Media	Basic Building Block	Slots Needed for Min/Max Configurations	Capacity (TB) Min/Max Configurations	Bandwidth (MB/s) for Min/Max Configurations
Fixed	StoreEngine (**3** drives)	1 to 4	3 to 12	750 to 2500

Removable	StoreEngine (**0** drives) + StorePak (**6** drives)	2 to 8	6 to 24	1400 to 5000

Fixed and Removable Media Options

The StoreEngine recorder provides options for both fixed and removable SATA storage media. The fixed media version is based on the StoreEngine VPX blade, each hosting three fixed SATA drives. The removable media version uses a combination of StoreEngine VPX blades, along with StorePak VPX removable SSD storage modules. Each removable StorePak module hosts six SATA drives, thus providing the same overall "drives per slot" density as the fixed media option. The fixed vs. removable storage options are illustrated in figures 12 and 13.

StorePaks support hot swap of SSD storage. A common usage model supported by StorePaks consists of conducting a mission, and at the conclusion of the mission, hot-swapping out the "full" StorePaks which contain the recorded data from the mission, and then hot-swapping in replacement "empty" StorePaks, making the platform immediately available for the next mission. The "full" StorePaks are then physically transferred to a post mission analysis facility where the recorded data is offloaded using one of the "playback" methods described earlier.

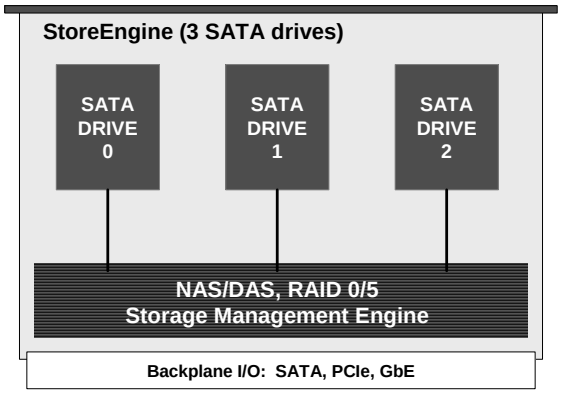

Figure 12. A StoreEngine blade configured with three non-removable SATA drives provides up to 1.5TB of storage in one 6U slot, and up to 6TB in four slots

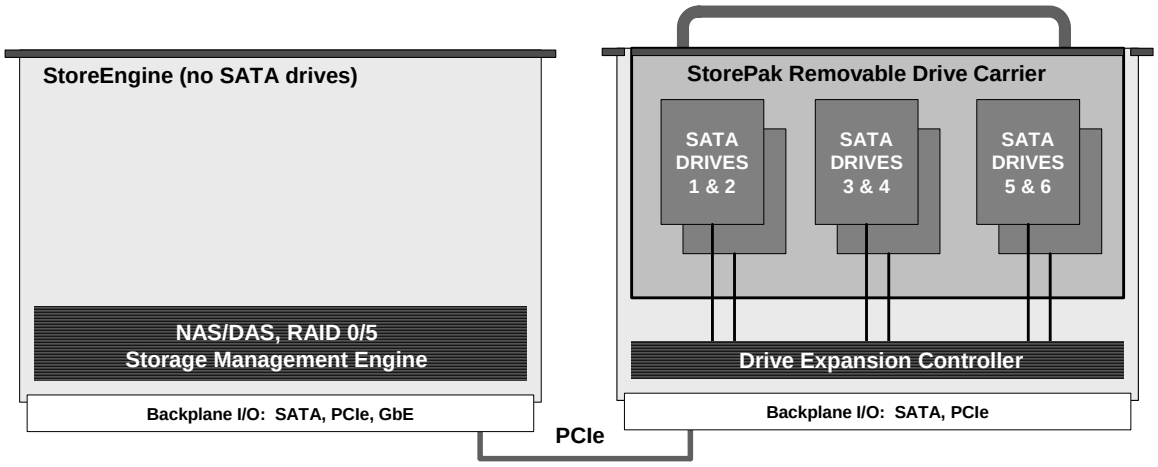

Figure 13. A drive-less StoreEngine blade, combined with one or more six-drive StorePak(s) blades, provides up to 3TB of *hot-swappable* SATA storage, using two 6U slots.