The TPT-RAID Architecture for Box-Fault Tolerant Storage Systems

Yitzhak Birk and Erez Zilber
The Technion – Israel Institute of Technology

***Abstract*—TPT-RAID is a multi-box RAID wherein each ECC group comprises at most one block from any given storage box, and can thus tolerate a box failure. It extends the idea of an out-of-band SAN controller into the RAID: data is sent directly between hosts and targets and among targets, and the RAID controller supervises ECC calculation by the targets. By preventing a communication bottleneck in the controller, excellent scalability is achieved while retaining the simplicity of centralized control. TPT-RAID, whose controller can be a software module within an out-of-band SAN controller, moreover conforms to a conventional switched network architecture, whereas an in-band RAID controller would either constitute a communication bottleneck or would have to also be a full-fledged router. The design is validated in an InfiniBand-based prototype using iSCSI and iSER, and required changes to relevant protocols are introduced.**

***Index Terms*—RAID, SAN, out-of-band, iSCSI, iSER, InfiniBand, RDMA.**

I. Introduction

In most current RAIDs [1], including very large ones, any given error-correcting (ECC) [1] group resides in a single box. Regardless of the degree of internal redundancy and reliability, a single-box RAID is thus susceptible to box-level failures (e.g., cable disconnection, flood, coffee spill), as these render entire ECC groups unavailable.

In a multi-box RAID, each ECC group uses at most one block from each storage box, so the failure of such a box does not render any data inaccessible. The controller must be fault tolerant (e.g., by having a hot backup [2]), as must the network [3]. Our work focuses on multi-box RAIDs with centralized control, and we use the term Multi-box RAID to refer to such systems.

Unlike a single-box RAID that uses a DMA engine for internal data transfers, a multi-box RAID must use the network, e.g., iSCSI over TCP, for all transfers. This requires extra data copies that affect both throughput and latency, and moreover burdens the CPUs. Overcoming the single point of storage-box (*"target"*) failure by going to a multi-box RAID thus poses several challenges: communication efficiency and prevention of a controller bottleneck. Controller fault tolerance can be handled through well-known mechanisms; it is not addressed in this paper, as the proposed architecture does not place any special demands in this respect.

[1]We focus on erasure correcting codes, mostly XOR, and use the term ECC loosely. Nonetheless, TPT-RAID can be adapted to use any ECC.

A. In-band vs. Out-of-band RAID Controller

Current SAN (block-oriented) controllers are either "in-band" (Fig. 1) or "out-of-band" (Fig. 2). In single-box RAIDs, the RAID controller is naturally in the data path. In a multi-box RAID, however, an in-band RAID controller is problematic:
– Connecting it to a single switch port renders it a communication bottleneck, as it is party to all communication.
– Connecting it via multiple ports may help but is costly, requires load balancing among the ports, and its internal data paths could be the bottleneck.
– Locating it inside the switch, acting as a router, would relieve the bottleneck, but the "orthogonality" of communication and other functions would be violated.

Fig. 1. In-Band controller

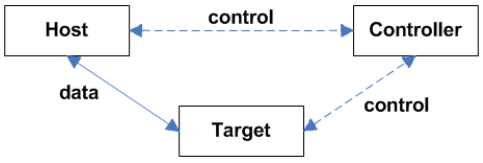

Fig. 2. Out-Of-Band controller

A multi-box RAID [2] comprising disk boxes and a controller, all interconnected by a common network, naturally admits an out-of-band RAID controller. However, this presents performance challenges and raises the issue of locating ECC calculations, which cannot be performed by such a controller.

B. Contributions of this work

We present the 3rd Party Transfer multi-box RAID architecture, TPT-RAID, which partitions the RAID controller functions: the management functions are taken out of the targets and placed in a centralized, out-of-band TPT-RAID controller, while data transfers and ECC calculations are carried out directly among targets and hosts and within targets, respectively, all under centralized control. The controller only handles control

[2]We use RAID-5 as an example, and refer to it simply as "RAID". However, this work is equally applicable to other RAID types.