



Thin provisioning with native hierarchical storage management

Author: Hewlett-Packard Company, Bob Cochran, Ayman Abouelwafa, Peter Djordjevich

Research Disclosure Database Number 531066

Published in July 2008

The Research Disclosure Journal is published on the 10th of every month. Every disclosure is recorded in the Research Disclosure database as soon as it is received and can be published online prior to being published in the next edition of the Journal.

Research Disclosure is a unique international defensive publication service that allows the world's intellectual property community to establish their inventions as prior art and prevent others from patenting the same. It is the world's longest running disclosure service. Due to the large number of Research Disclosure citations appearing in patents world wide, WIPO has recognised Research Disclosure's importance by granting it PCT Minimum Documentation Status.

Kenneth Mason Publications Ltd give consent for this disclosure to be printed out providing it is for personal use, or for the personal or internal use of patent examiners or specific clients only. Photocopies may be made providing it is for personal use, or for the personal or internal use of patent examiners or specific clients and not for resale and the copier pays the usual photocopying fee/s to the relevant Copyright Clearance Centre. This consent does not extend to abstracting for general distribution for advertising, or promotional purposes, for creating new collective works or for resale. This consent also does not extend to other kinds of scanning, printing or copying, such as printing, scanning or copying for general distribution for advertising, or promotional purposes, for creating new collective works or for resale. Document delivery services are expressly forbidden from scanning, printing or copying any Research Disclosure content for re-sale unless specifically licensed to do so by the publishers.

Research Disclosure Journal, ISSN 0374-4353

© Kenneth Mason Publications Ltd

www.researchdisclosure.com

info@researchdisclosure.com

Thin provisioning with native hierarchical storage management

High Level Introduction:

Today, storage array based Thin Provisioning is limited to constructing virtual volumes from paged pools made up of a single, homogeneous storage type. Since ~90% of a typical application LUN is dormant on any given day, this results in a high level of cost/power/cooling inefficiency. For instance, even if only ~10% of a Mission Critical application LUN is highly utilized on a given day, 100% of that LUN will typically reside on the most power hungry, hottest running and expensive type of disk spindle (e.g. 15kRPM FC). This application moves the state of the art forward by bringing what is historically known as host-based HSM into the storage controller to periodically and automatically migrate individual ThP pool pages and LUSE (aggregated) LUN constituent LDEVs (Logical Devices) to performance appropriate storage to greatly improve storage cost efficiency. The result is cost efficient Storage Array LUN storage utilization for both LUSE and paged pool virtualized volumes (e.g. ThP [Thin Provisioning]).

What if:

- LUSE volumes could consist of thousands of LDEVs
- *ThP pages can be any multiple of 42MB in size. E.g. 4200MB*
- ThP pools had the option to be exclusively a collection of (large) **page sized** Open-V LDEVs (e.g. each pool LDEV=16800MB, which is a multiple of today's 42MB default page size)
- The concept of a ThP 'Pool Group' existed
- A ThP volume is allowed to consist of pages from different pools within its designated pool group
- Each ThP volume is assigned a 'default pool' within its pool group
- The array FW (firmware) tracked the daily average activity level of each LDEV in a LUSE and each allocated page in a ThP volume (new)
- AutoLUN (internal data migrator software [SW]) worked with individual:
 - ThP pages, pools and pool groups
 - the LDEVs of a LUSE
- Any pool group could auto-expand any of its pools from equivalent storage RAID groups [RGs] in a 'reserved RG pool', from which authorized pool groups and LUSE LUNs can draw

General Procedure

Define the possible storage types hierarchy and the cut line for acceptable high availability, based on the various possible combinations of factors such as:

- Spindle RPM (e.g. 15,10,7.2 kRPM) for both internal and or vs. external disks
- External 3-9s availability, Vs 4-9s, Vs 5-9s, single path'ed Vs Multi-path'ed, path fail-over, or not, load balanced, or not
- SATA with or without 'read verify after every write' enabled
- Power/cooling requirements (100% for FC, 90% for SATA)
- Solid state storage, Etc.

Procedure for ThP

- Create the individual pools for a particular ThP pool group. E.g. ‘ThP-PG1’ =
 - I. 15kRPM FC internal (consisting of 1000 LDEVs of 16800MB each= ~16TB)
 - II. 10kRPM FC internal (consisting of 2000 LDEVs of 16800MB each= ~32TB)
 - III. 7.2kRPM SATA internal (consisting of 4000 LDEVs of 16800MB each= ~64TB)

- At ThP volume definition time, designate:
 - the pool group name and page size for that ThP vol. (e.g. ‘ThP-PG1’, 16800MB page size)
 - the ‘default’ pool for that ThP vol. (e.g. #2 above – middle of the road for performance, power & cooling)
 - (optionally) Some reserved & unallocated RAID Groups making up the ‘Reserved RG pool’, which can be used to create Open-V LDEVs of the necessary size to augment any compatible pool with a nearly full status
 - (optional) Permission to migrate¹ pages to another pool if the currently hosting pool is:
 - not the optimal storage type for the activity level of that page
 - dangerously full and the other (default demotion to slower) pool has adequate space

ThP Example – day 1

All allocated pages originally derive from the medium cost/performance pool (top diagram at right).

ThP Example – day 2

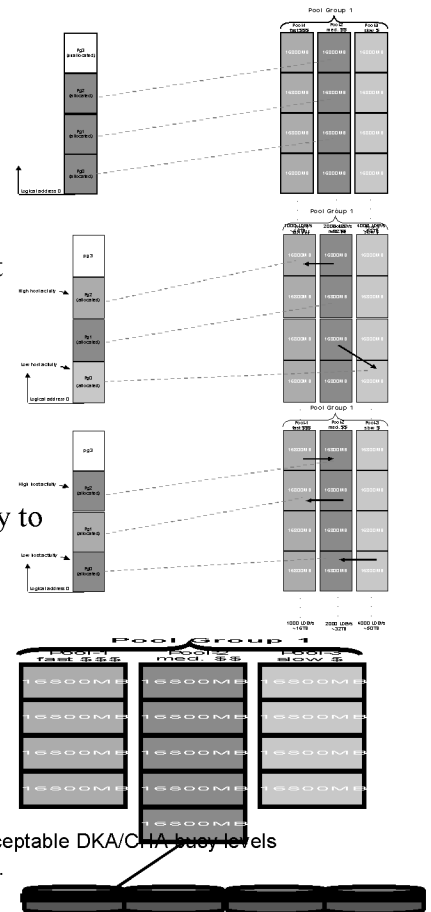
One page shows very high host activity and another shows very low host activity, so they’re automatically migrated via AutoLUN’s auto-migration feature to more performance appropriate storage.

ThP Example – day 3

Based on the last activity sampling period (e.g. 24hrs), pages are again automatically promoted or demoted by maximum one pool step at a time. The algorithm is weighted so an allocated page will have an affinity to stay with its assigned default pool (e.g. pool 2).

ThP Example – day 4

Due to (on demand) ThP page allocations, Pool 2 is nearly full, so the array automatically creates an appropriately sized LDEV from a compatible reserved RAID Group (e.g. 10kRPM internal) and appends



¹ Typical AutoLUN algorithms for activity sampling rates/periods, automatic data migration and acceptable DKA/CHA busy levels will be applied to promote or demote a page to the next faster or slower pool from the approved list.

the LDEV to the pool. The reserved RAID Group is available for use by multiple Pool Groups. The array also has the option to demote some of the lower activity pool-2 pages to the next level down pool (if capacity exists). Promotions and demotions only occur if space is available.

Procedure for LUSE

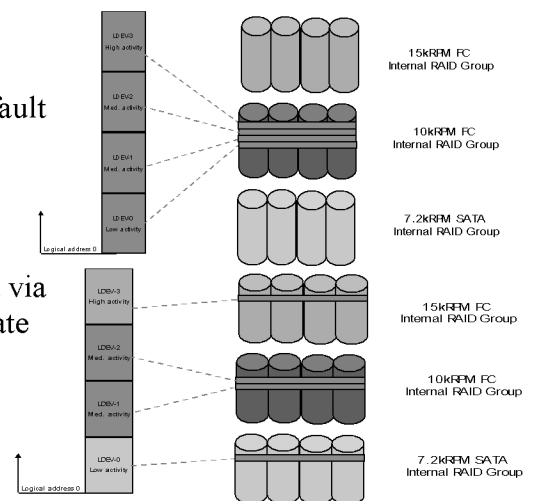
At LUSE definition time: identify the 'RAID Group Pool' and the constituent RAID groups (e.g. the 3 below) that are allowed to participate in the automatic creation, migration and deletion of LUSE constituent LDEVs. Identify the default RG number (and type) for that LUSE (e.g. #2 below). RG 1-1, 15kRPM FC internal 3D+1P, RG 2-2, 10kRPM FC internal 3D+1P & RG 3-3, 7.2kRPM SATA internal 6D+2P

LUSE Example – day 1

All the LDEVs making up a LUSE derive from its assigned default RAID Group.

LUSE Example – day 2

One LUSE LDEV shows very high host activity and another shows very low host activity, so they're automatically migrated via AutoLUN's (prior art) auto-migration feature to more appropriate storage.



Summary of Advantages for THP

- I. Instead of 90% of every THP volume's pages being on the wrong storage type/speed/cost for any given day (because only 10% of a LUN is typically changed in a day), XP ThP-HSM reduces it to < ~20%.
- II. Storage Administrators can add capacity to the reserved pool with much less urgency as compared to prior art (instead of waiting for a more urgent pool threshold alert).
- III. Allows customers to skew the internal drive type tier-ing percentages (e.g. more SATA than normal) for improved power/cooling without any application aware performance penalty.

Summary of Advantages for LUSE

- I. Instead of 90% of all LUSE LDEVs being on the wrong storage type, speed, cost power/cooling for any given day (only ~10% of a LUN will be modified in a day), this method reduces that to < ~20%.
- II. Allows customers to skew the internal drive type tier-ing percentages (e.g. more SATA than normal) for improved power/cooling without any perceived application performance penalty.

Note: Given the examples illustrated above involving LUSE and ThP of a standardized page size of 42MB, it may become obvious to an implementer that the performance accommodation algorithm(s) can be the same for both products and potentially even more array features. This page size standardization (if elected) could potentially be applied to snapshot, Auto-LUN, CA-J performances issues as well.

Disclosed by Bob Cochran, Ayman Abouelwafa, Peter Djordjevich, Hewlett-Packard Company