

HITACHI DYNAMIC TIERING WEBTECH SERIES SESSION 1 OF 3

**STEVE BURR, SOLUTION ARCHITECT, GSS SERVICES
ENGINEERING**

**JOHN HARKER, SENIOR PRODUCT MARKETING
MANAGER**

JULY 13, 20 AND 27, 2011



Hitachi Dynamic Tiering WebTech Series

Learn about fine-grain automated storage tiering, which used to be an exotic technology available from only a few sources. Now it is rapidly becoming a standard capability of advanced storage systems. To use it properly, you need to understand new factors in system design, such as an application's storage locality of reference pattern. In this webinar series, you will hear from the experts about how to take best advantage of this new technology on Hitachi Virtual Storage Platform. The 3 sessions in this webinar series will take you through useful topics with prepared remarks and interactive discussion:

- A deep look at the theory and design of Hitachi Dynamic Tiering
- How to size and configure a system
- Installation and operation

You'll learn how to:

- Determine if an application's data is a good fit for automated tiered storage movements.
- Select and size tiers within a Dynamic Tiering pool.
- Use replication and migration with Dynamic Tiering virtual volumes.
- Properly operate and monitor a Dynamic Tiering system.

HITACHI VIRTUAL STORAGE PLATFORM

HITACHI DYNAMIC TIERING WORKSHOP

PART 1 – PRINCIPLES
PART 2 – CONFIGURATION
PART 3 – OPERATION

STEVE BURR – SOLUTIONS ARCHITECT, GLOBAL SOLUTIONS SERVICES

JOHN HARKER – SENIOR PRODUCT MARKETING MANAGER

JULY 2011

- A workshop and masterclass series on Hitachi Dynamic Tiering
- Summarize and contrast with Hitachi Dynamic Provisioning, if appropriate
- Technical in nature (where it aids understanding)
- Content-heavy
- To include many recommended practices
- Some new material
- Aims to help your evolution with the technology and therefore maximize the benefits
- Three parts
 - Part 1 – Principals
 - Part 2 – Configuration and Design
 - Part 3 – Operation

- Introduction
- Classic and Dynamic Provisioning
- Dynamic Tiering
- Monitoring
- Relocation
- Tier Range and Performance Optimization
- Benchmarks

INTRODUCTION

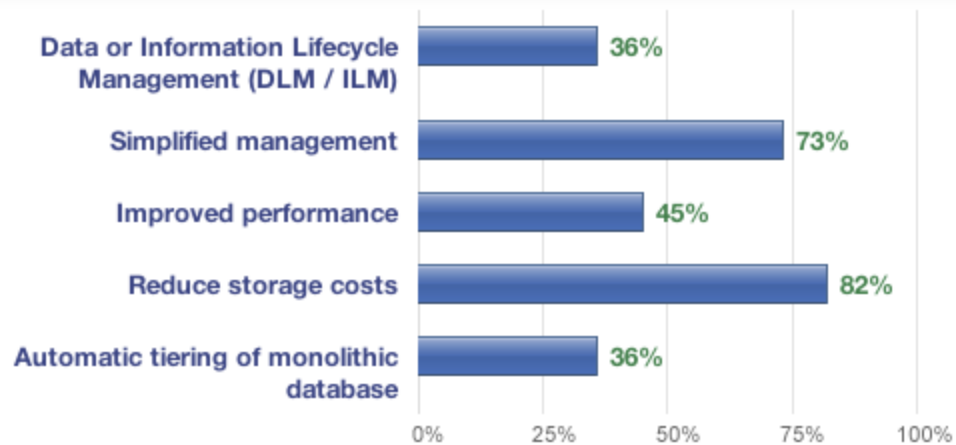
HITACHI
Inspire the Next **Dynamic Tiering Customer Feedback**

“ It removes the complexity of managing a multi-tier environment. ”

Source:  System Administrator, Global 500 Insurance Company

 TechValidate
TVID: D81-DBF-9FD

Why did you decide to use Hitachi Dynamic Tiering virtual volumes? Choose all that apply.



Note: this is a multiple-choice question – response percentages may not add up to 100.

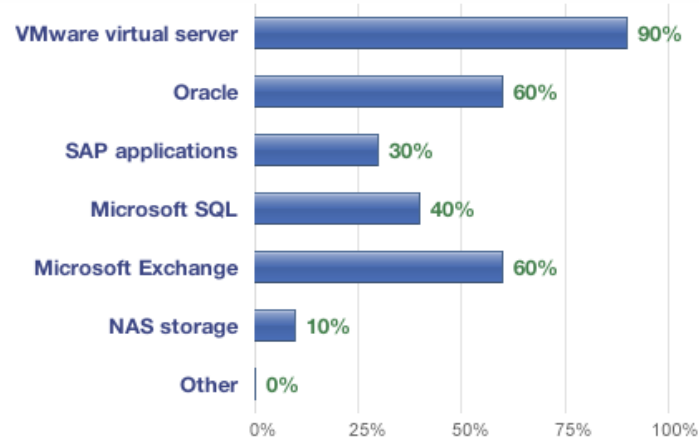
Source:  TechValidate Survey of 11 Hitachi Virtual Storage Platform Users

 TechValidate

TVID: 757-1EC-403

DATE: 03/20/11

What applications do you use with Dynamic Tiering? Choose all that apply.



Note: this is a multiple-choice question – response percentages may not add up to 100.

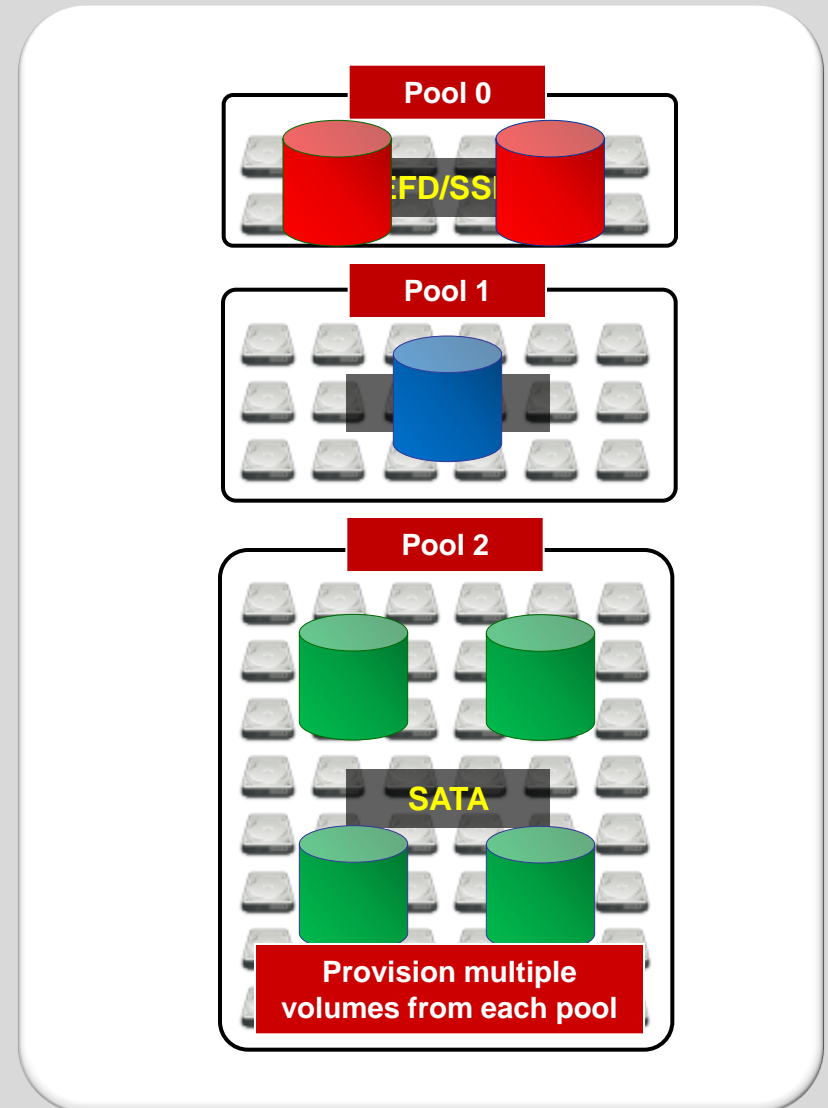
Source:  TechValidate Survey of 10 Hitachi Virtual Storage Platform Users


TVID: B44-161-C78
DATE: 03/20/11

CLASSIC PROVISIONING
DYNAMIC PROVISIONING

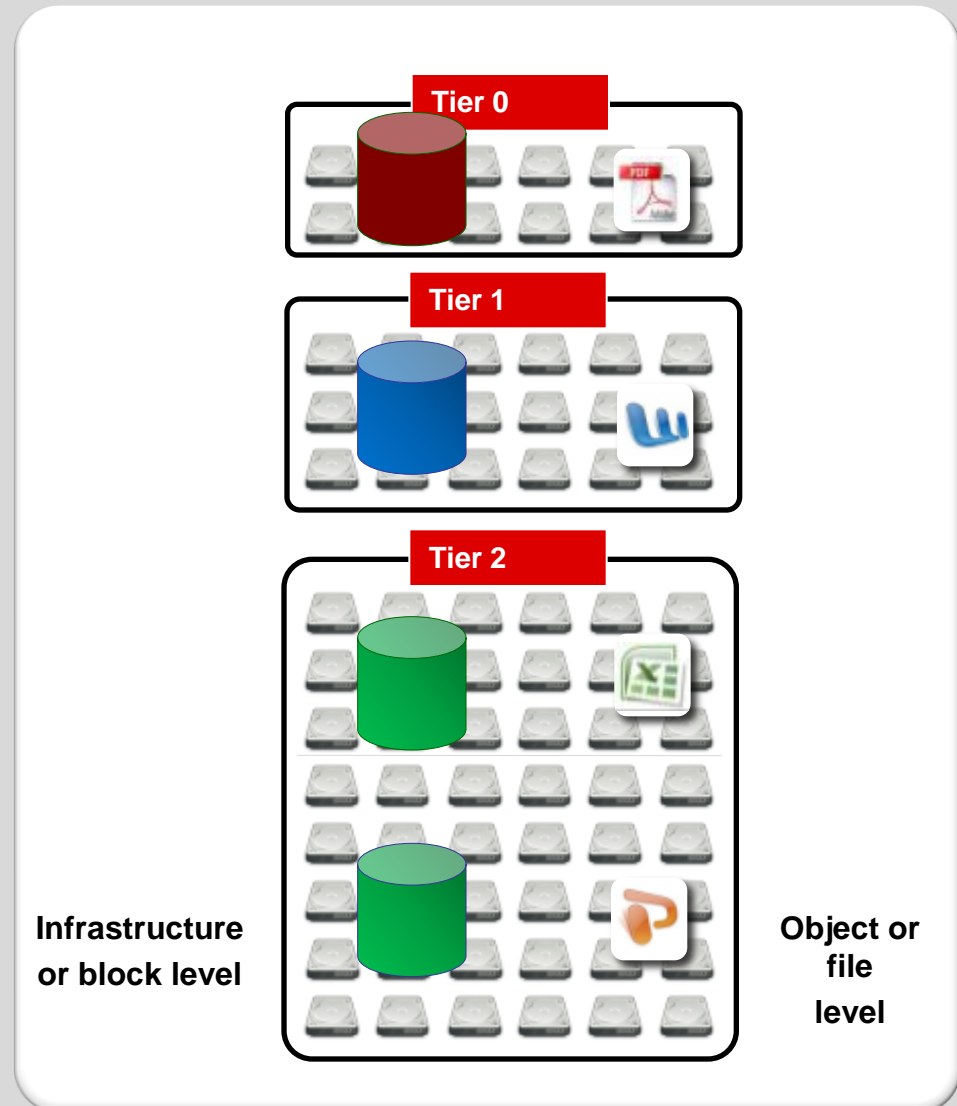
CLASSIC PROVISIONING WITH STORAGE TIERS

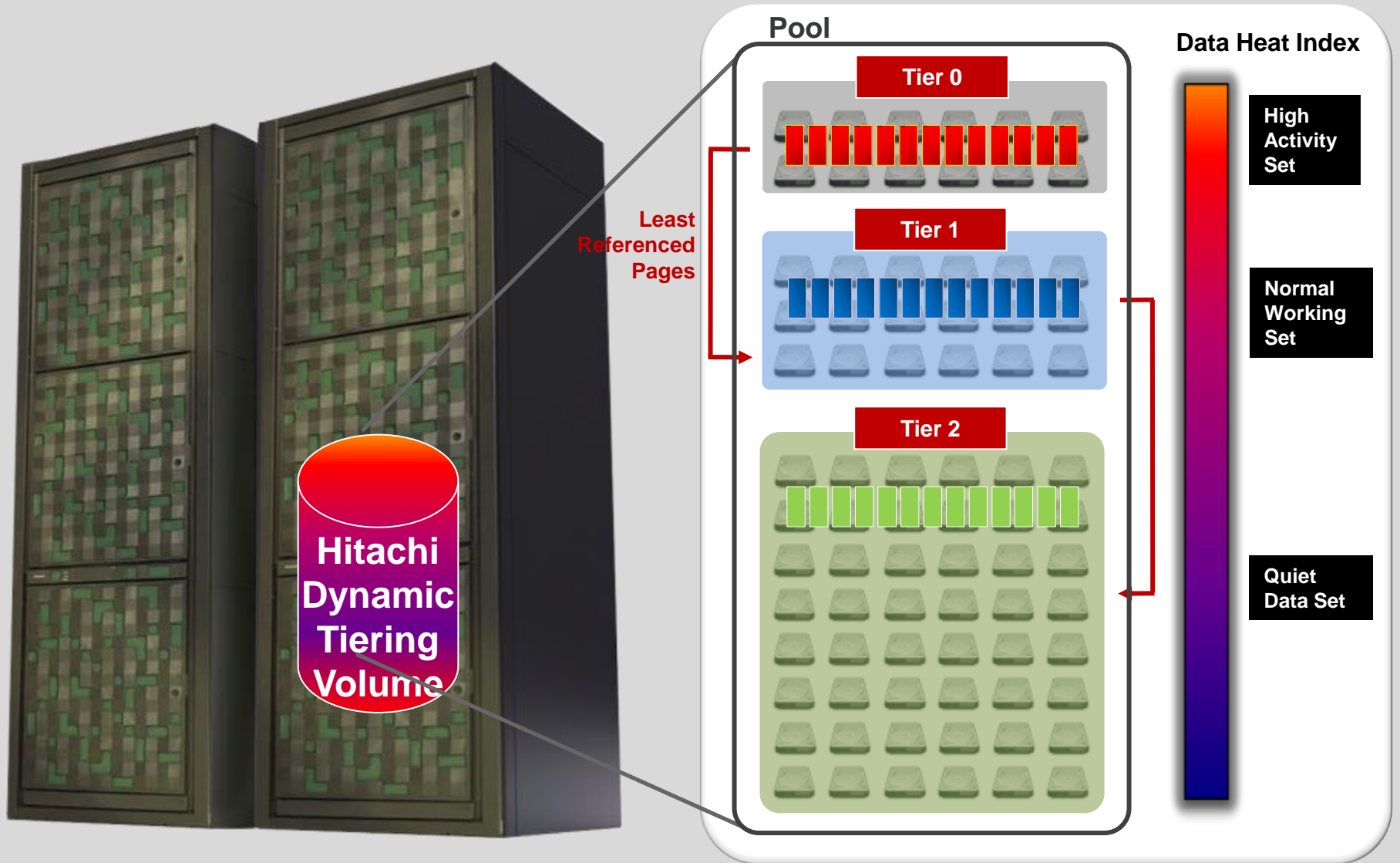
- **Start with drive technology**
 - A simple criterion that gives us a sense of cost and performance
- **Define tiers**
 - SSD = Tier 0
 - SAS or FC drives = Tier 1
 - SATA drives = Tier 2
- **Advanced: Use Hitachi Dynamic Provisioning and define separate pools**
 - Ease of provisioning
 - CAPEX deferral and avoidance
 - Performance improvements



CLASSIC DATA MOBILITY

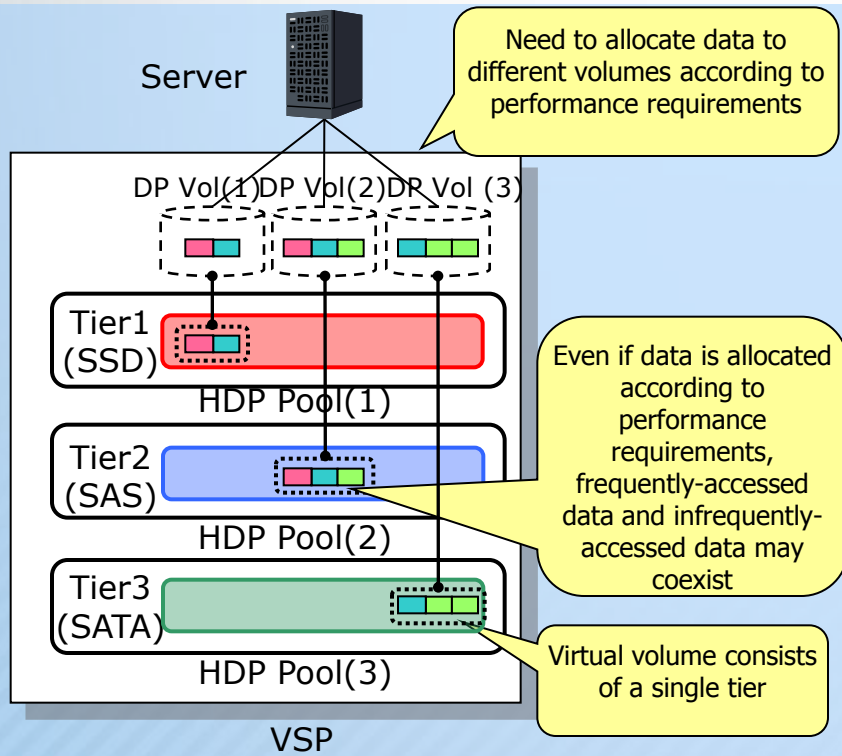
- **Two layers of data mobility**
 - Infrastructure layer or block level
 - Object layer or file level
- **Use virtualization and data mobility tools to move volumes without disruption to another pool or tier**
 - Promotion or demotion
 - Consolidation or migration
 - SLO, performance or cost change
- **Automated with policy-based management**
 - Based upon pre-set SLAs



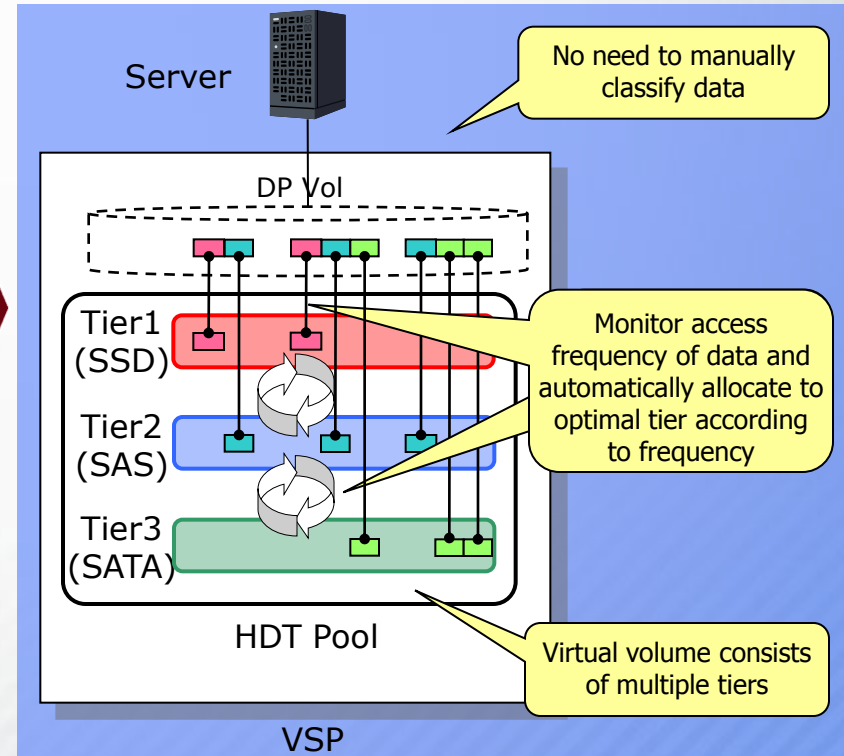


- Combines multiple media, each with different cost performance (SSD, SAS, SATA)
- Relocates highly accessed data to high-speed media and infrequently accessed data to low-speed media according to access frequency
- Monitors I/O load per page, much smaller than a volume, and allocates pages to the optimal media tier

Before (Hierarchical Control per Volume)



Hitachi Dynamic Tiering



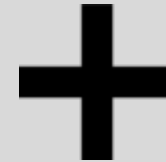
■ : Frequently accessed data
 ■ : Data with moderate access frequency
 ■ : Infrequently accessed data



Dynamic Tiering

**Dynamic
Provisioning**

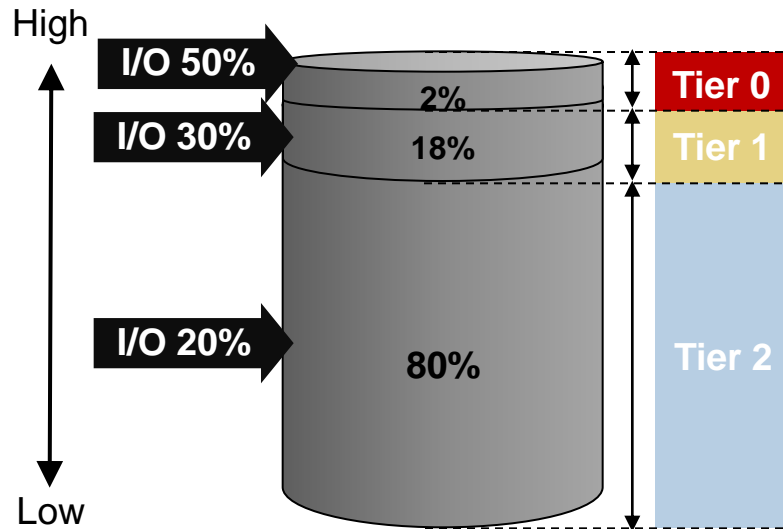
- All the benefits of Hitachi Dynamic Provisioning



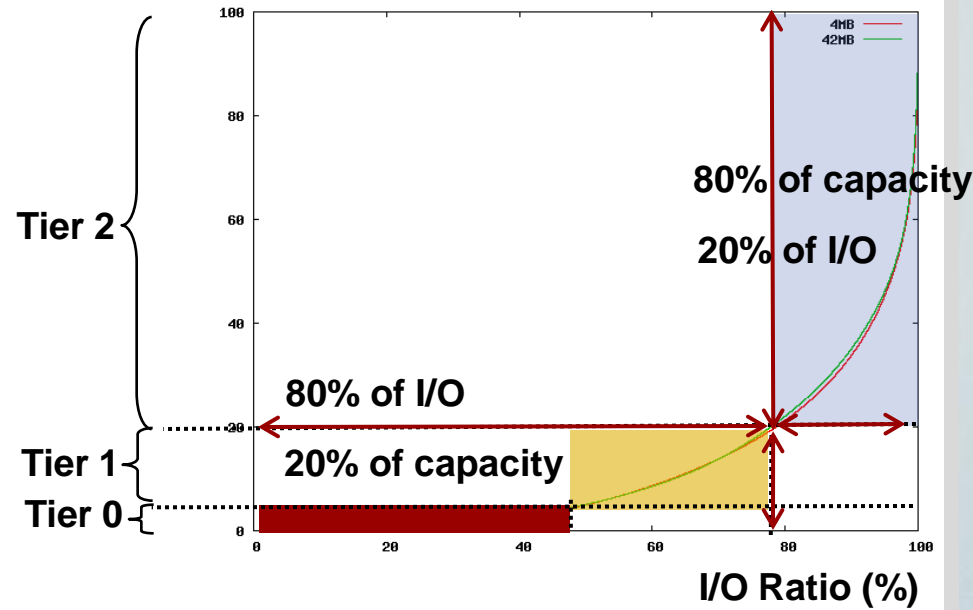
- Further simplified management
- Further reduced OPEX
- Further improved return on assets (ROA)

IMPROVED PERFORMANCE AT REDUCED COST HOW IT WORKS – THE 80/20 RULE

Workload



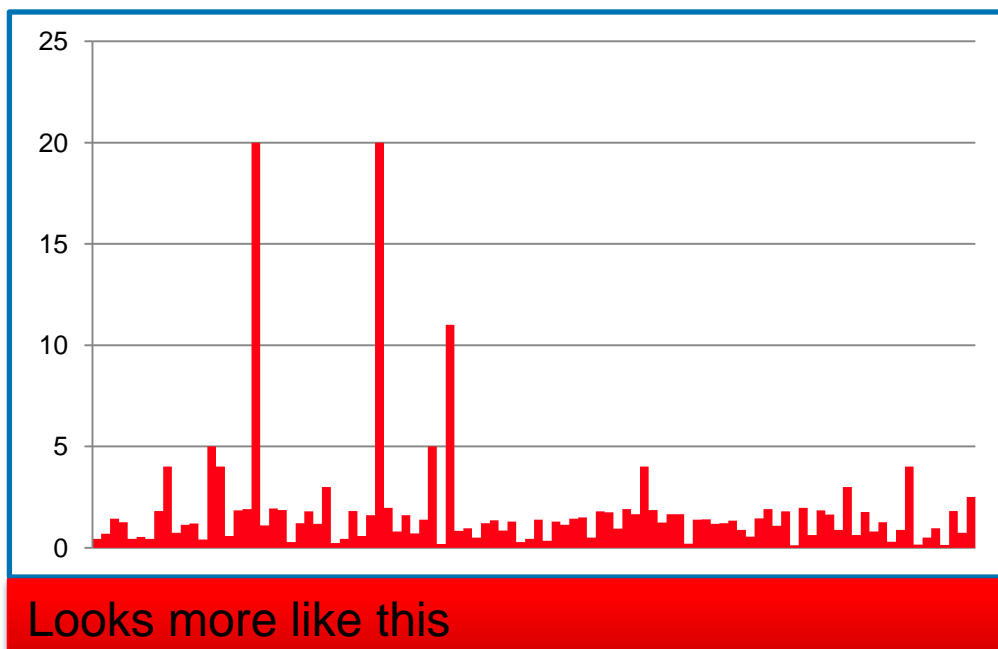
Capacity Ratio (%)



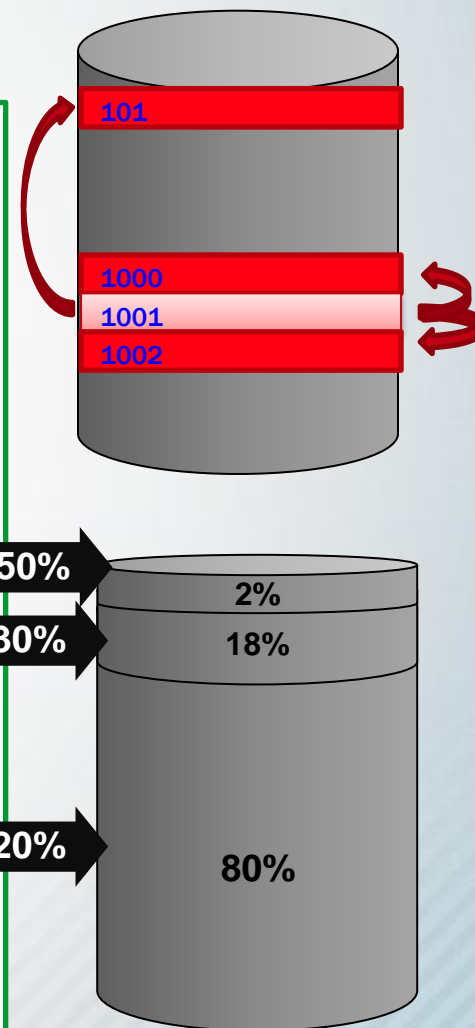
Research and Investigation

- Analyzed access locality of volumes for one year
- 80% of I/O concentrated in 20% of the total area
- 50% of I/O concentrated in 5% of the total area

- What makes Hitachi Dynamic Tiering work



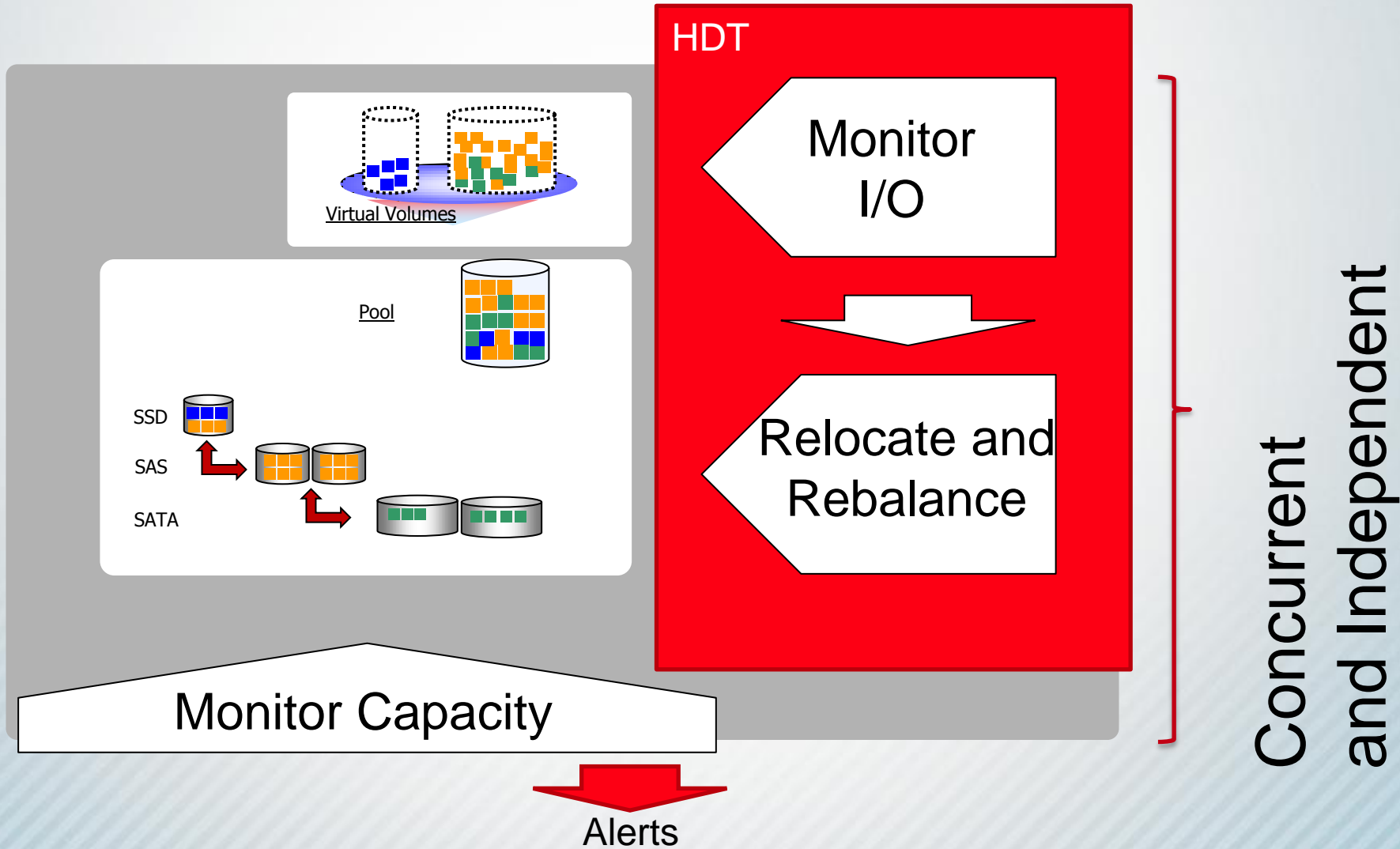
We show it like this for clarity



LOCALITY OF REFERENCE

- Temporal locality of reference: With real applications, we find that accessed data is close to other data that will be accessed in the near future.
This makes cache work.
- Spatial locality of reference: With real applications, we find that frequently accessed data is often close to other frequently accessed data, and rarely accessed data is often surrounded by other rarely accessed data.
This makes Hitachi Dynamic Tiering work.

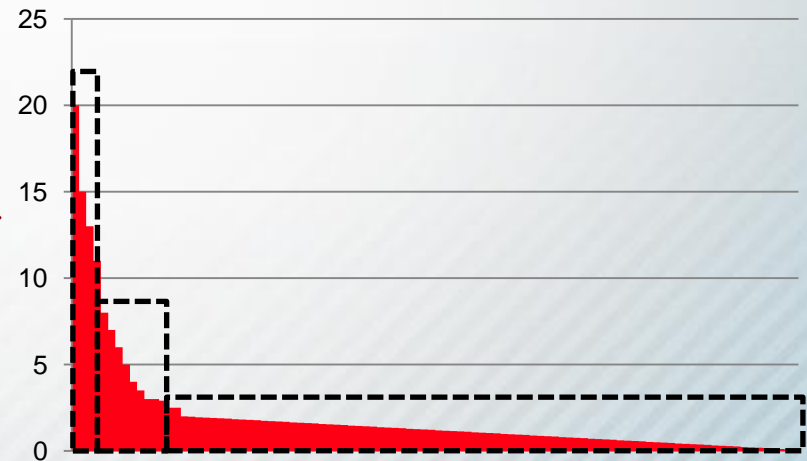
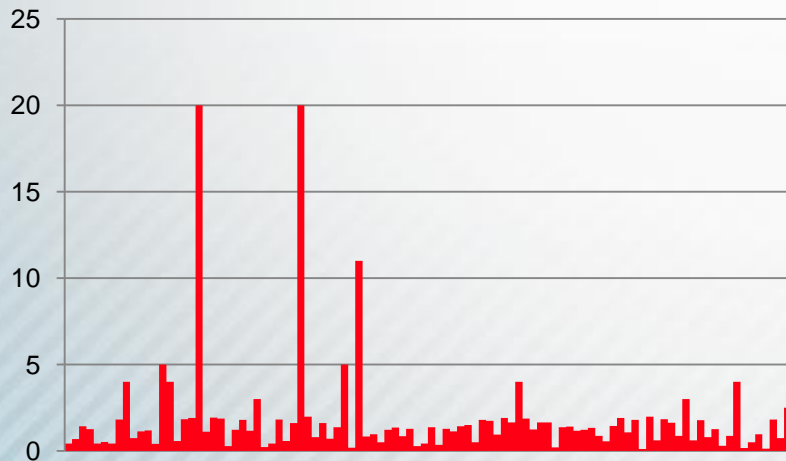
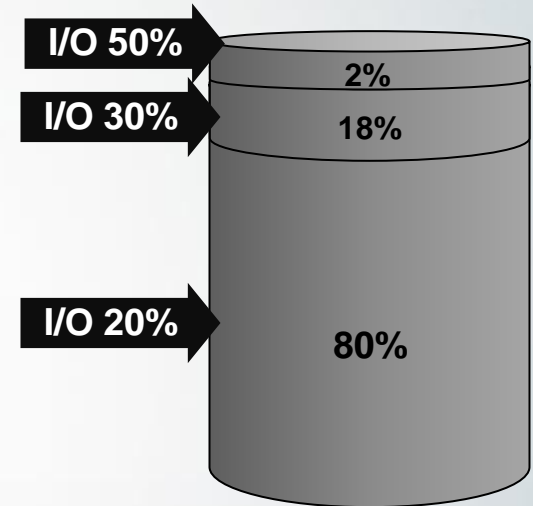
THE MONITOR RELOCATE CYCLE



MONITORING

HDT PERFORMANCE MONITORING

- Each page has back end I/O counted (read+write) during the monitor period.
- Locked pages are not counted.
- Some pages under management are ignored.
- Pages, not LUNs
- We order to create a distribution function.
 - IOPH vs. GB
- This is analyzed to decide the “best” tier for each page.



- Manual mode
 - Monitoring and relocation separately controlled by either script (raidcom) or Hitachi Storage Navigator Modular 2
 - With scripting, can set complex schedules

- Automatic mode
 - Customer defines strategy, and it is executed automatically
 - Period or continuous monitoring
 - 24 hours
 - Defined part of 24 hours
 - ½, 1, 2, 4 or 8 hourly
 - All aligned to midnight

MONITORING OPTIONS

Execution mode	Execution cycle	Performance monitoring		Relocation		Monitoring and relocation cycle
		Start	End	Start	End	
Auto execution	24 hours [monitoring time not specified]	After setting Auto execution to ON, next 0:00 is reached	After monitoring started, the next 0:00 is reached	Start immediately after monitoring info is fixed	One of the following <ul style="list-style-type: none"> Relocation of entire pool is completed Next relocation is started Auto execution is set to OFF 	
	24 hours [monitoring time specified]	After setting Auto execution to ON, the specified start time is reached	The specified end time is reached	↑	↑	<p>[Ex.] Monitoring period 9:00-17:00</p>
	30min, 1h, 2h, 4h, 8h	After setting auto execution to ON, cycle time starting at 0:00 is reached	After monitoring started, cycle time is reached	↑	↑	<p>[Ex.] Monitoring period 8h</p>
Manual execution	-	Request to start monitoring is received	Request to end monitoring is received	Request to start relocation is received	One of the following <ul style="list-style-type: none"> Relocation of entire pool ended Request to stop relocation is received Auto execution is set to ON 	

(*) Time zone is Storage Local Time. Summer time not supported

PERIOD AND CONTINUOUS MONITORING

Period mode

Monitor current cycle only

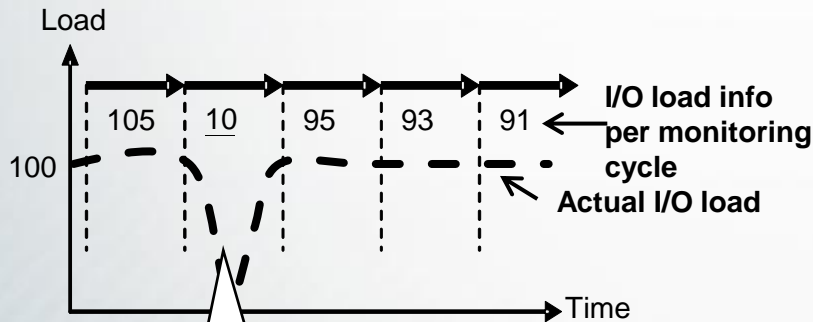
Relocation executed according to the current I/O load

Continuous mode

Monitoring repeated cycles. Weighted average of current and previous cycles used.

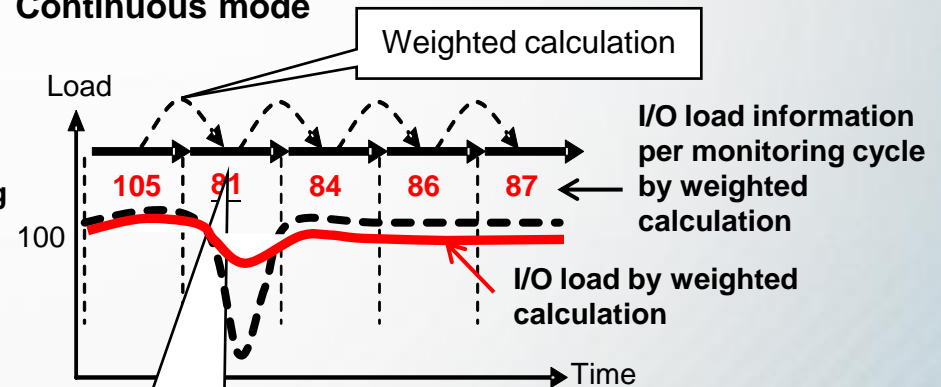
Relocation due to short-term increase or decrease in I/O load avoided.

Period mode



Relocation executed based on current I/O load

Continuous mode



Relocation executed based on weighted calculation

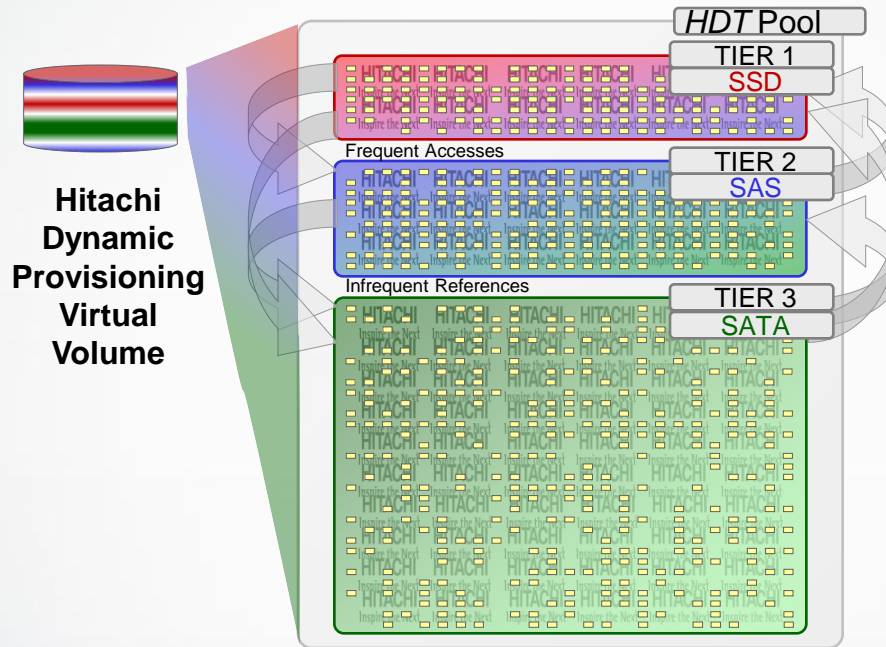
- Short monitor and relocate cycles were too responsive to temporary changes for some customers.
- With period monitoring, the only solution was a longer cycle.
- Combining continuous monitoring and a relatively short cycle may give best of both worlds.
 - Fast promote: A rise in workload that extends over multiple cycles is likely to be promoted
 - Slow demote: A drop in workload will not be actioned unless extended over multiple cycles

Continuous mode – more information:

- Information from previous days, weeks, even months is included in the weighted number.
- Histogram and other displays are weighted data only.

- A short automatic cycle will be more reactive, but may not be long enough to relocate everything.
- **Priority is upward relocation**
- A long automatic cycle will keep pages stable but may not promote “fast enough.”
- Do not expect to rapidly swap in and swap out diverse work patterns.
- For best results, size higher tiers to keep all of the frequently accessed pages for the day (or longer).
- Manual mode can provide very long relocation periods and custom patterns.
- **For weekly or daily patterns, consider continuous mode.**

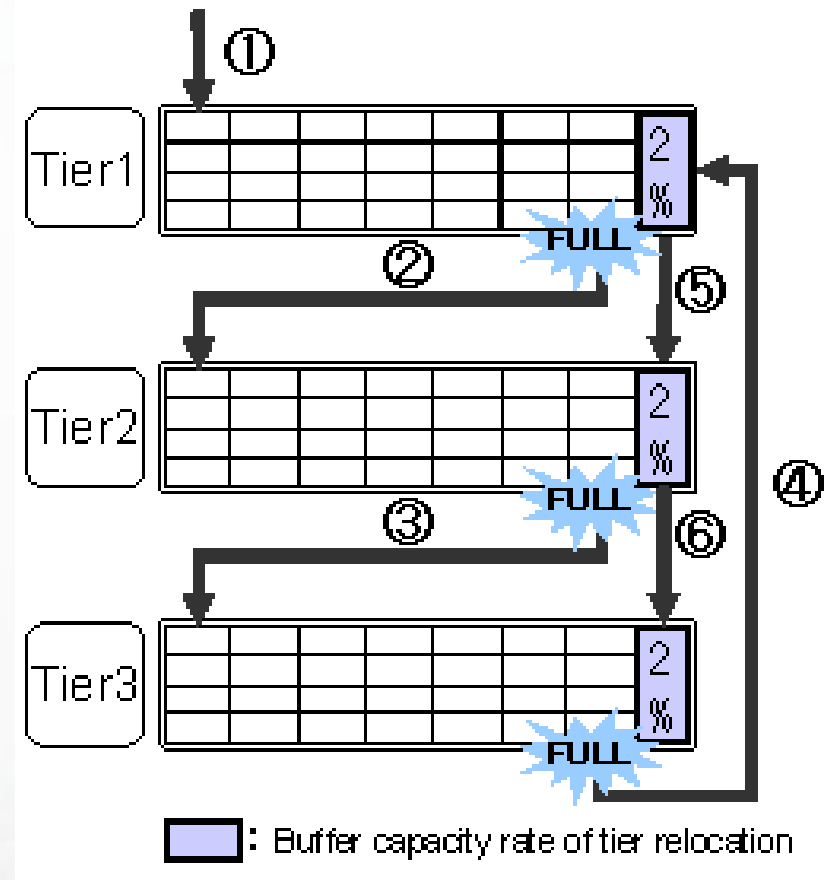
RELOCATION



- What determines if a page moves up or down?
- When does the relocation happen?
- Caution: To aid understanding, earlier slides in this presentation simplify

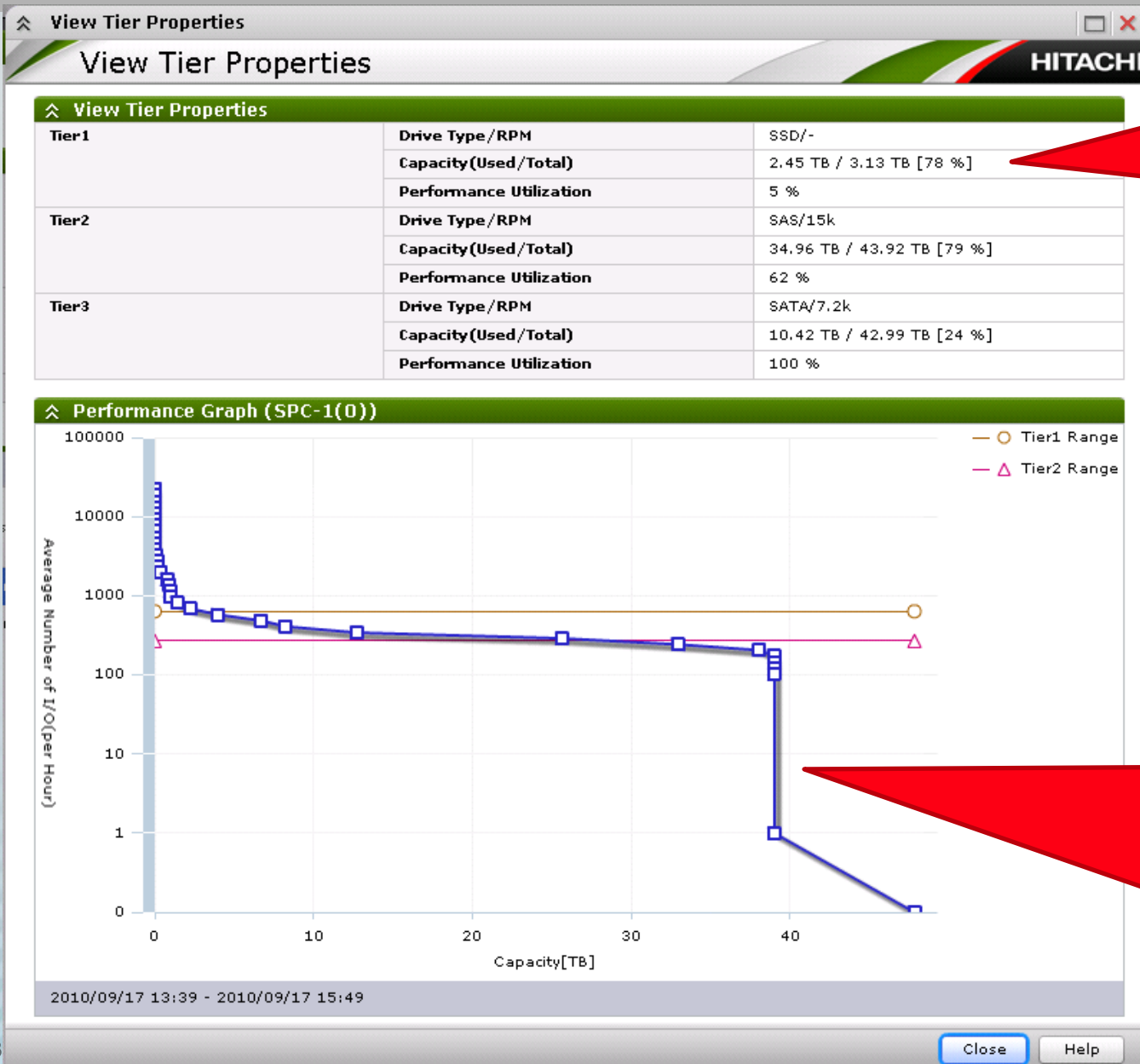
- Two aspects, handled differently
 - New pages
 - Relocated pages
 - Space has to be reserved for both
- The default reservations are:
 - R% reserved for relocation: Variable 2-40% per pool, default 2%
 - T% reserved for new pages: Variable 0-50% per tier
 - Solid state drive (SSD) default 0%
 - Other drive types default 8%
- Changing these numbers is rarely recommended, because it can reverse the intended effect.

- Allocate to highest tier with free capacity
- This keeps best resources busy
 - However, a new pool is not representative for early users.
 - Playing with T% is not the answer.
 - If concerned, build the pool slowly.
- SSD defaults to 0% reserve
 - Most new page requests in a mature pool with SSD go to Tier 2.
- Later, you will see that there is another factor to change this.



- At the end of a monitor cycle, we have a list of pages sorted from hottest to coolest.
- This is used for the decision-making of the next cycle.
- Monitoring has dual buffers, so monitor and relocate overlap.
- The goal is:
 - Hottest should be in the highest tier
 - Coolest should be in the lowest tier
- You can see this in “pool tier properties.”
This is incredibly powerful and has many uses.

POOL TIER PROPERTIES



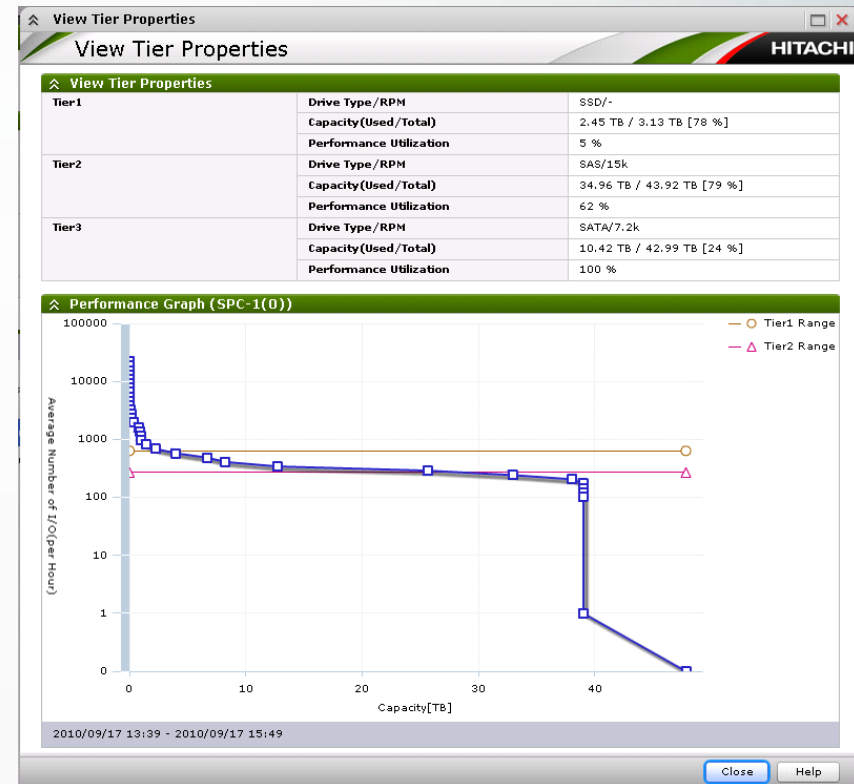
What is being used now in the pool in terms of capacity and performance

The I/O distribution across all pages in the pool

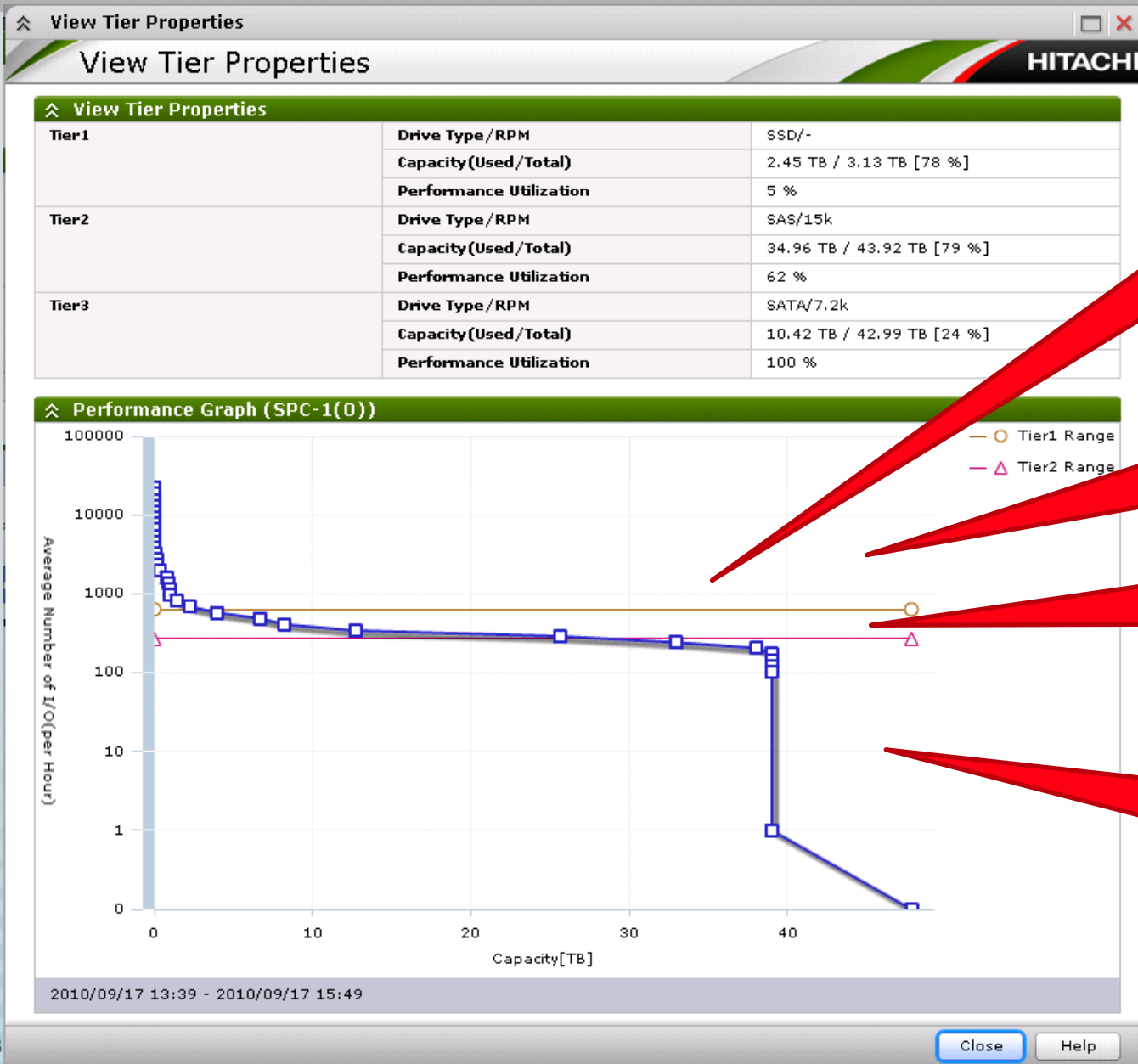
Combined with tier range, this is what HDT is using to decide where the pages should go to.

Things you may want to know

- IOPH scale is similar to logarithmic
- Not continuous, but separate, points (lines help viewing only)
- Points above tier range 1 are not “pages in T1”, they are pages that ideally should be in T1.
- The relocation algorithm will try to relocate pages to their correct tier.
- Relocation may not complete in shorter cycles.
- New monitor data will then be used.
- Tier range moves!
- The shape of the curve tells you how well Hitachi Dynamic Tiering will perform.



TIER RANGE: RELOCATION DECISIONS



Data from the previous cycle

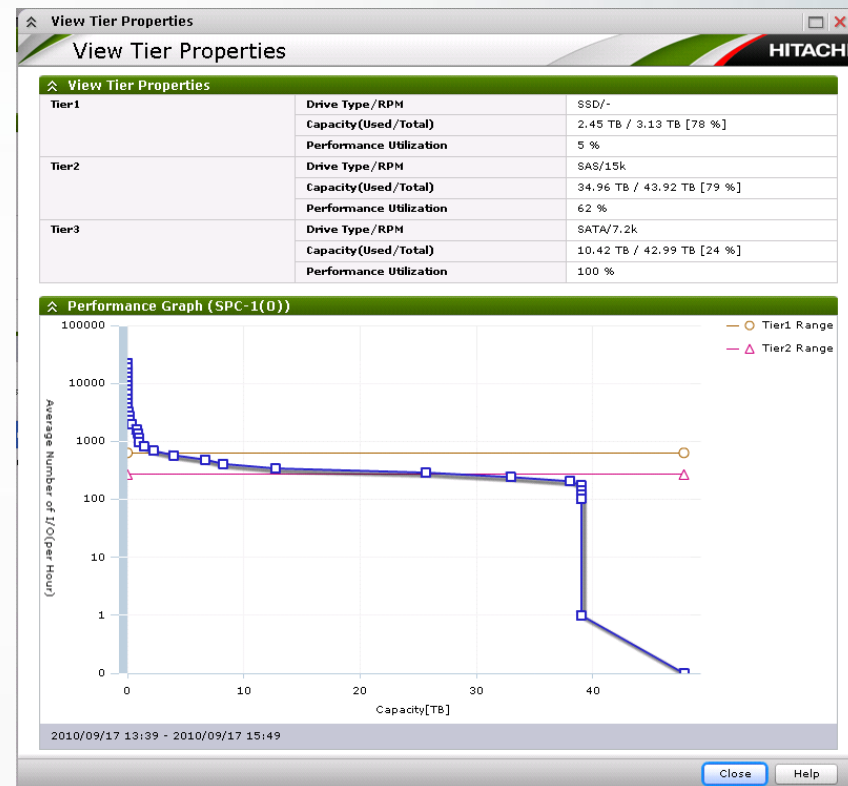
Pages above tier range 1 are candidates to be in Tier 1.

Pages above tier range 2 are candidates to be in Tier 2.

Pages below tier range 2 are candidates to be in Tier 3.

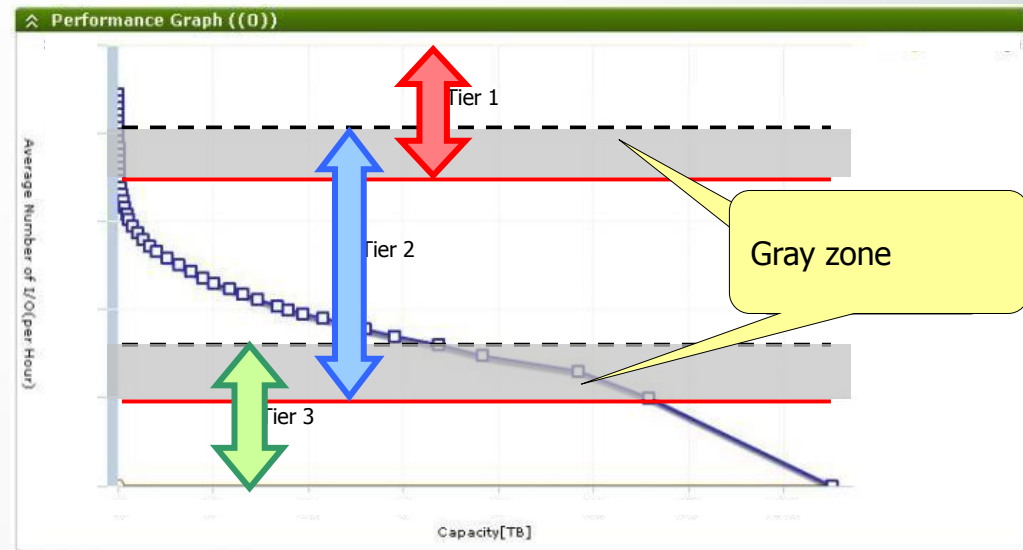
TIER RANGE: GRAY ZONE

- You might think that tier range is just the IOPH value such that each tier is completely filled.
- We first adjust reserve space for relocation and new pages.
- Pages vary in IOPH every cycle.
- It would be wrong to move a page too near a tier range as it could be moved back next cycle.
- We widen the tier range to give a “gray zone.”
- Hover over the tier range to see.



GRAY ZONE AND RELOCATION DETAILS

- To avoid pages bouncing in and out of a tier, data in the gray zone is left where it is, unless the difference is two tiers.
- Does relocation impact production?
 - Relocation is 3TB per day (35MB)
 - Calculated to not impact production
 - Similar to rebalance
- A free zero page reclamation (ZPR) is done during relocation



HDT OPTIMIZES CAPACITY AND PERFORMANCE

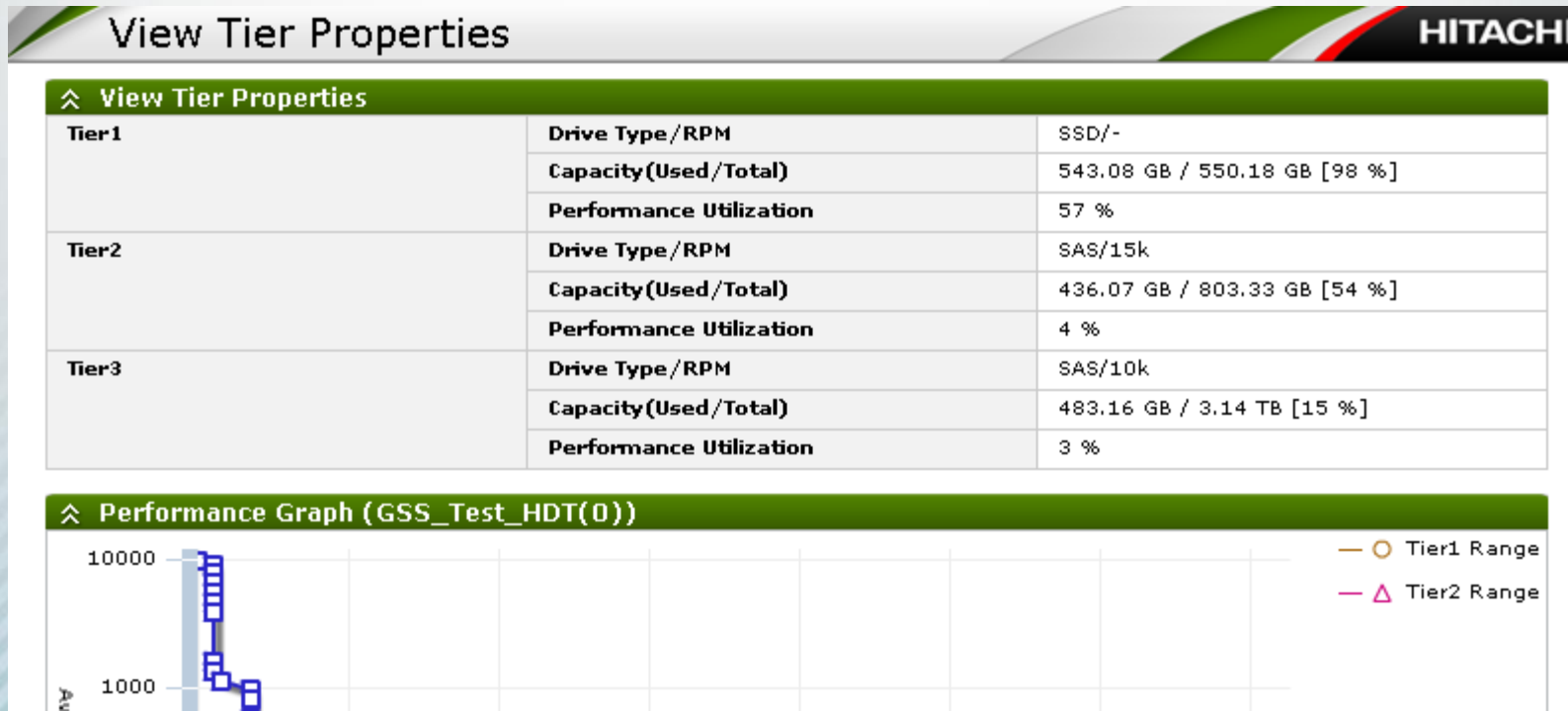
- The above description is fine while each tier is operated within capacity **and** performance capabilities.
- **You should aim to operate HDT pools so all tiers are running within their design parameters.**
- But what if performance is near or over the achievable throughput?
 - Total IOPS too high: whole pool suffers
 - Peak IOPS too high: overload Tier 1
 - Any tier can be overloaded
- We cannot prevent a user making too many requests to a pool but we can **try to do the best with what is available.**
 - As T1 approaches “ideal operating level,” we stop using it.
 - As T2 approaches “ideal operating level,” we stop using it.
 - Should T3 also be overloaded, we start to “share the pain.”
 - Overload the pool in all tiers – the pool is overloaded

HOW WE DO THIS

- We tune tier range dynamically.
- If a tier approaches 60% utilization, we aim to move I/O down a tier.
 - 60% sustained I/O to accommodate peaks and prevent queuing
 - Tier range is reduced. This reduces the tier's capacity. All of your tier will not be used, but this is better than overloading it.
- If all tiers are over (60%,60%,60%), the pool is overused. Tier ranges are increased again to share the problem.
- You can't lose storage, but you might put more I/O into a pool than it can handle.
- This is **not** Hitachi Dynamic Tiering going wrong. It is trying its best – better than I could do.
- You should be looking for these situations, evaluating the issues and adding capacity.

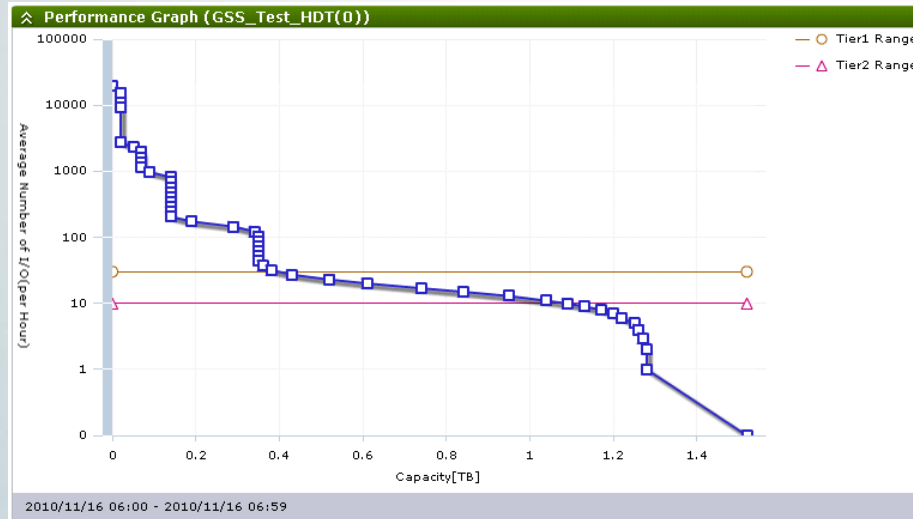
HOW WE MODEL TIER PERFORMANCE

- Not dynamic measures but IOPH counts vs. performance modeling
- Each drive has a nominal design performance per gigabyte
- Tier performance capability as capacity multiplied by nominal performance per gigabyte
- This is calculated dynamically and shown in tier properties as **performance utilization** or **P%** in **raidcom get dp_pool -key opt**

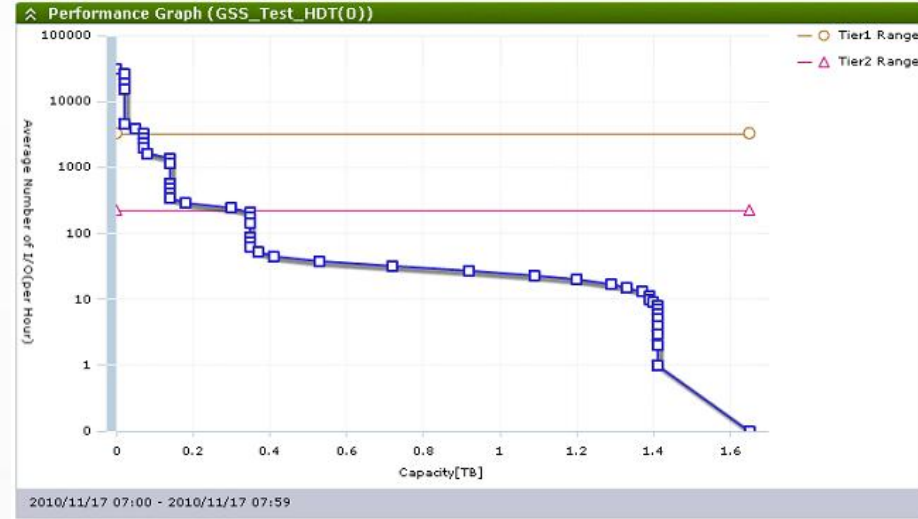


TIER RANGE MOVEMENT IN PRACTICE

5,000 IOPS



7,000 IOPS



PID	POLS	MODE	STS	DAT	TNO	TL_RANGE	TD_RANGE	TU_CAP(MB)	TT_CAP(MB)	T(%)	P(%)	R(%)
000	POLN	AUT	RLM	VAL	1	0000001e	00000008	549066	563388	0	57	76
000	POLN	AUT	RLM	VAL	2	0000000a	00000006	646884	822612	8	2	76
000	POLN	AUT	RLM	VAL	3	00000000	00000000	545748	3293808	8	0	76

PID	POLS	MODE	STS	DAT	TNO	TL_RANGE	TD_RANGE	TU_CAP(MB)	TT_CAP(MB)	T(%)	P(%)	R(%)
000	POLN	AUT	RLM	VAL	1	00000ccd	000002de	111300	563388	0	59	99
000	POLN	AUT	RLM	VAL	2	000000e2	00000038	186606	822612	8	57	99
000	POLN	AUT	RLM	VAL	3	00000000	00000000	1442868	3293808	8	14	99

Note: Tier performance utilization alone isn't a good signifier.
Relative tier capacity used says more.

BENCHMARKS: SYNTHETIC

- This is not “how to do it” but “why you probably won’t succeed.”
- It is very hard to create a benchmark like real life.
- This can be hard to explain, since many aspects are new.
- Some benchmarks will perform unrealistically fast.
- Some benchmarks will perform unrealistically slow.
- Most benchmarks don’t have representative spatial locality of reference.
- Look at the tier properties performance graph in Storage Navigator Modular 2.
- You will have to create much larger datasets than usual.
 - Benchmark data must fill Tier 1 and Tier 2 and some of Tier 3.
- Soak tests are useless.
- HDT performs a free zero page reclamation (ZPR) when migrating between tiers. We’ve seen customer benchmark data reclaimed completely (Most synthetic benchmarks send empty blocks).

BENCHMARKS: APPLICATION

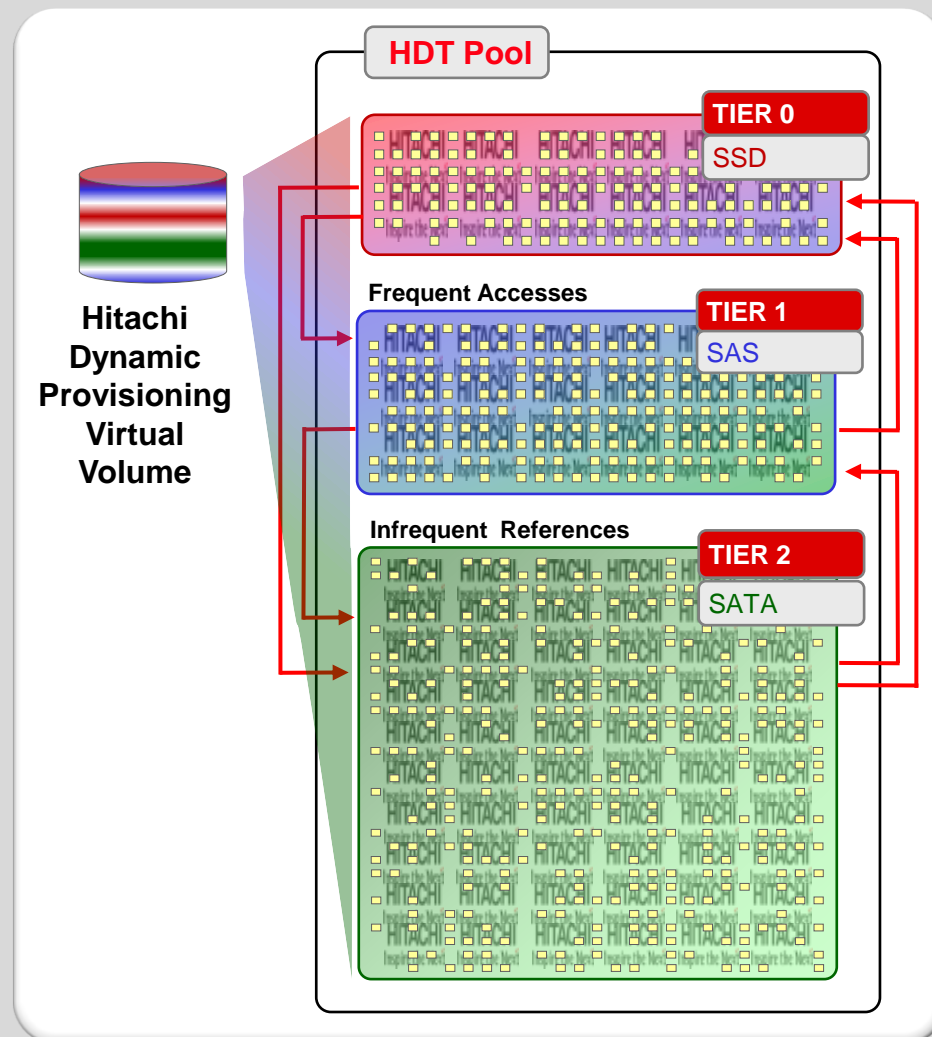
- Only the real application in a real pool with all its planned real workloads is truly representative.
- And then it's only representative of that time of test (Add more applications, and the workload is spread).
- Synthetic application benchmarks may have the correct read/write and random/sequential profile, but will they have correct record locality? Very possibly not.
- Workload profile simulators are not likely to be good.
- Even the real application won't be 100% if the workload is synthetic.
- Example: Database with representative but synthetic transactions
 - Probably representative of log and database I/O, read/write, etc.
 - May even be representative of table locality.
 - But will the table I/O have correct locality of customer access?

Before: Tiered Storage and Provisioning

- Front end scripting
- Data classification before tiering
- Complicated management of multiple storage tiers
- Different SLAs for different tiers

Now: Dynamic Tiering and Provisioning

- Controller-based automation
- Single, self-managed, self-healing, efficient pool of data
- All the benefits of tiered storage
- All the benefits of Hitachi Dynamic Provisioning
- No need for data classification



- **Simplifies operations and data management**
- **Reduces OPEX, CAPEX and TCO**

QUESTIONS AND DISCUSSION

July

- Hitachi Dynamic Tiering WebTech Series (3 Sessions), July 13, 20 and 27 at 9am PT, 12pm ET (1 hour each Wednesday for 3 weeks)

August

- Ensure Data Relevance with Hitachi Data Discovery, August 3, 9am PT, 12pm ET
- Save Power. Save Space. Save Money, August 17, 9am PT, 12pm ET

Please check www.hds.com/webtech for:

- Link to the recording, the presentation and Q&A (available next week)
- Schedule and registration for upcoming WebTech sessions

THANK YOU