List of Pending Patents at USPTO: 2018 - 2020

Rahul Vishwakarma

Areas: Data Storage, Persistent Memory, Machine Learning, Blockchain

1 Method to suggest best SCM configuration based on resource proportionality in a de-duplication based backup storage

Storage class memory (SCM) supports operation in memory, App-Direct and storage modes. We can convert SCM to perform in these modes by changing different parameters such as sector size, mapping, alignment (2k, 4k), devdax, fsdax and size of namespace. If we have SCM in IO path then data passes through different layers such as PMEM aware filesystem, BTT (Block Translation Layer) driver, MMU mapping and hardware layer. Each layer adds latency for considerable amount of time. Some layer has characteristics of working better in certain configuration (read type, write type, block size). All these layers are NOT involved in all kind of configuration. SCM supported block sizes are 512, 520, 528, 4096, 4160 4228, where it supports SSD attributes.

- All these supported blocks have their own merit and de-merit along with performance. All the writes are not persistence in nature. These data streams will be kept in different caches between CPU and memory. CLFLUSH, CLWB, CLFLUSHOPT instructions are introduced to make all the writes persistence
- These commands must be used optimally for better system performance.

2 Distributed Architecture for Thin Provisioned Storage Class Memory

It can be difficult to size how much Storage Class Memory is needed and where it is best deployed and adding it individually to servers can result in underutilization without enough applications being designed to take advantage of it. Proposed Solution:

- JBOSS provides a way to consolidate SCM investment and provision it throughput the enterprise with RDMA fabric
- SCM can be run in hybrid mode where it can be used as RAM until needed for persistent memory

3 Accelerating Backup by placing synchronous write IOs to Performance Ranked SCM in RDMA environment

When we have more than one SCM installed in the server then the performance of one SCM will differ from other one. If we select better performing SCM for

synchronous write, then overall write performance will increase. Same performance will reflect when read of data will happen from same SCM. Until system configuration (Hardware configuration) is not changed, we can leverage this idea to boost the overall performance. Memory to memory copy happens very fast. This technique can be used to boost restore performance (same technique can be used in cluster environment). When Storage Class Memory (SCM) is exposed through Remote Direct Memory Access (RDMA), a portion of SCM memory region can be leveraged by different RDMA Initiator(s). When multiple SCM from different servers in a clustered environment are exposed then the users can choose the best performing SCM to the application(s) becomes one of the most crucial and deciding factor. This proposed solution will benefit data stream coming from source and destination to SCM location, which will indirectly improve overall IO performance; furthermore, this inherently improves the total customer experience.

- Seamless RDMA experience leveraging storage class memory
- Dynamic management of SCM namespace (Portion of memory)
- Using "Global View". Auto assignment of apps with server using
- Heterogeneous index along with "Access latency".

4 State semantics kexec based firmware update

Today, one of the prime value additions for customer is reduced downtime during an event such as 'upgrade', which eventually leads to shorter upgrade time. To address this issue, we have already implemented 'kexec' based upgrade in Data Domain.

In kexec process we load the second (new) kernel first. Internally it allocates new pages for second kernel. Along with new pages kexec also creates a control page of the size of PAGE SIZE. Later after "kexec-execution" second (new) kernel alone with initrd is booted.

Proposed solution implements Extensible-Control-Pages (ECP) in kexec (Control pages are the intermediator between old kernel and new kernel). Extensible-Control-Pages will have the information has to be passed to new kernel and they are elastic in nature. When kexec transitions to the new kernel there is a one-to-one mapping between physical and virtual addresses. In that environment kexec copies the new kernel to its final resting place.

In new solution, along with kexec-load new pages are allocated in the multiple of PAGE SIZE for user data. This memory buffer will be used by user to copy custom data. After kexec load Extended-Control-Pages are added to existing control-page link list. This is achieved by having a hook in "copy user segment list" call. Once second (new) kernel is booted Extended-Control-Pages are extracted and passed to required application.

5 System and method for an efficient dual-relationship based hash structure for Non-Volatile Memory technology

In the field of persistent memories, non-volatile memories (NVM) have become quite promising as a replacement for traditional DRAM technology owing to their non-volatile nature, byte-addressability, low power consumption and high storage density. However, since they have a different hardware architecture altogether, the memory structures built for DRAM become inefficient if used in NVMs. Hashing index structures, which are fundamental data structures to provide fast response to queries is one such case. Moreover, the writes in NVM are costly, hence there is a need for the hash structure has to be write-optimized. Here, we are proposing a new hashing data structure that is efficient over the existing structures in terms of –

- optimized writes by shared buckets
- search domains
- computational cost by using a unique single hash fucntion
- update operation by conditional logging
- resizability: rehashing reduced to a factor of – 0.14 in expansion and 0.57 in shrink

6 Boost restore bandwidth and reduce latency by ACPI assertive method

In a deduplication-based backup system, when we perform backup using Data Domain Boost – the backup operation (using distributed segment processing) is faster as compared to our competitors; however, the restore operations are still a pain point for customers which takes longer than the backup operations. The reason is primarily because of the bad locality of data on backup storage disks due to deduplication mechanism. Although, we have tried to reduce the restore operation time using to be at par with customer's expectation, we still see a room for improvement.

Here we propose a method for accelerating the restore performance in a deduplication-based backup system based on ACPI assisted operations by:

- Matching memory initiator (restore job) and memory target (heterogenous domain) path with best available memory bandwidth
- We rank the heterogenous domain based on access, read bandwidth and read cache attribute of memory (alter allocating restore job to these virtual domains)

7 Method for securely archiving digital information in oligonucleotides

Digital information storage in DNA is not a new idea and the recent development in synthesis of DNA has significantly reduced the price of per base pair (DNA-Based unit of storage); however, the current price per base pair remains USD 0.07, which is far from commercial implementation as compared to HDD or SSD based storage cost. Prior work are more focused on designing a data encoding method for DNA-Based storage, error correction code, genomic compression, enzymatic method of synthesizing nucleotides in lab, and few works on encrypting the data. To best of our knowledge, very less work has been done in reducing the amount of nucleotide required for digital information representation and finally reducing the storage cost per unit of storage for real time use. Few of the noticeable problems from in-silico perspective can be stated as below:

- Efficient encoding of data type (text, image, binary files) and random access
- Choice of deduplication method for genomic data as compression doesn't help much
- Mechanism for eliminating single point of failure in tradition DNA-Based storage
- Encoded data in DNA is not resistant to malicious tampering

We propose a method whose goal is to address the above problems and subsequently reduce the storage cost for DNA-Based storage per unit of digital information. We achieve this objective as follows:

- Random Access – Key Value Pair for Locality Similarity Hashing (LSH)
- Less number of nucleotides for information representation – Similarity based variable length deduplication with delta encoding for genomic data (at destination)
- Elimination of single point of failure – Storing metadata in Blockchain which makes data immutable and tamper proof

8 Stochastic risk scoring with counterfactual analysis for storage capacity

Storage management is one of the most discussed topics across the storage vendors because of its crucial importance to business. Even in a finely-honed backup regime, there are situations of DU/DL because the overall estimation process does not evolve over time with respect to changing workloads. Furthermore, safeguarding high accuracy of capacity forecast estimates along with ease of

interpretability and recommendation to the user plays an important role for any customer facing tool. Today, most of the machine learning and statistical methods used for storage management only provides a point in time when the storage is going to be full, and in few cases it also provided recommendation for adding extra storage or automated ordering of disk drives; however, in this disclosure we shift the focus from forecasting a single estimate for date of attaining full capacity to predicting the risk associated with running out of storage capacity. Furthermore, we also fine-tune the recommendation system which can explain the cause and effect of system behavior over data growth using model-agnostic counterfactual analysis – for example, our proposal can recommend administrator to check data movement or data retention policy.

- We address the challenge of dynamic non-linear chaotic time series data by using Geometric Brownian motion with drift. The volatility (drift) and shock (variance) of the model is derived by weighted causal features from explainable SHAP scores
- Our results show that a probabilistic approach is more accurate and credible, for systems with non-linear patterns, compared to a regression (Isilon), segmented-regression (Data Domain) or ensemble forecasting models CloudIQ

9 Discerning scrub based on multi-probabilistic approach

Disk scrubbing is a process of performing full media pack sweeps across allocated and unallocated disks and if latent medium error is detected then rebuild the missing data, which in turn reduces the chance of bad block media detection during host IO activity. However, running scrubbing task for entire population of disks in an array significantly increases the load of the data storage system, and may degrade its performance. There are lot of storage solutions where they focus on correcting the data, and scrub function is periodically called to scrub a portion of a component's on-disk data structures. These storage solutions bring the scrub functionality to much higher level (application level) and create minute scrubbing threads which are scheduled for a fraction of time. The work performed by the scrub routine is a relatively small portion of work (on the order of 1sec or so). Scrub functions for various components are called in sequence in order to interleave scrubbing. If a single scrub routine takes an inordinate amount of time, then it can starve the other scrub routines.

To overcome the impact from this issue in any Server and minimize the burden of scrub, we can only scrub the disk for which the operation is really required. We identify this using a learning framework based on multi-probabilistic approach to pin-point specific disks for scrubbing. The method is machine learning agnostic and translates to a binary classification task where we proactively forecast (n-days ahead) the state of a disk into two categories:

- Drives having more likely to have issues (we called them 'concern' drives)

- Drives with no issue (we call them 'no-concern' drives)

We create a set of above categories and quantify "how much" concern/no-concern is it across the entire storage pool based on prediction's confidence. This metric is used to prioritize (selective) scrubbing for the drives. The proposed method has two-fold advantages. First, the method can be used to forecast disk drive failures. Second, the quantified output from the forecast engine can be fed as an input for Scrubbing scheduler engine.

10 A data placement method based on node's health score in dedup cluster

Data placement in dedup cluster is an important topic in industry. Existing methods mostly focus on capacity balance, workload balance or global dedup ratio optimization. As we know, few of them take each node's health into account when deciding to distribute the backup data. In this IDF, we propose a data placement method based on the cluster node's health status. Firstly, we collect each node's runtime statistic including storage, CPU, memory, network and other HW platform factors. And by the trained model, it evaluates node's health score based on the collected data. Secondly, to align with customers' SLA agreement like system down time or RTO (recovery time object), we assign different customers' backup data onto nodes which have different health scores. Our solution's principle includes: A. critical backup business use the best healthy nodes; B. backup which can tolerate temporary down time use medium healthy nodes. C. proactively offline and diagnostic poor healthy nodes. We highlight the follows as protect points for filing a patent:

- Node health score method based on conformal prediction framework
- Backup data placement strategy based on nodes' health score

11 Extended list of pending invention disclosure at USPTO 2018 - 2020

- Intelligent Data Restorability using Algorithmic Randomness Model
- Method of selective disk scrubbing based on algorithmic randomness
- Variable Sparing in Large RAID Groups Based on Disk Failure Probability Analysis
- Context-aware intelligent maintenance window identification
- Approximating replication completion time based on algorithmic randomness
- Survival forecasting of disk drives using semi-parametric transfer learning

- Probabilistically Forecasting Health of Hardware in a large-scale system
- Explainable leaning model-based prioritizing and preventing backup failures
- Estimating replication completion time
- Dynamic Hot Sparing based on Ranking of Failed Disk Drives in Large-Scale Storage Infrastructure
- Analyzing Time Series for Large Populations of IoT Devices
- Autonomous and Dynamic Resource Allocation in Storage Systems
- Efficient backup system aware direct data migration between clouds
- Drive failure prediction based on incremental learning
- Optimized and Intelligent Data-movement in a Backup Storage Systems
- Reliable Health Forecasting of Enterprise Disk Drives
- Avoiding backup failure based on cold data
- Autonomous resource setting for garbage collection in storage systems
- Avoiding backup failure in a deduplication-based storage systems
- Dynamic backup policy handshakes based on capacity projection
- Capacity forecasting in backup systems