Systems Research and Innovation in Data ONTAP

Scott Dawkins
Vice President
Advanced Technology Group
NetApp Inc.
Scott.Dawkins@netapp.com

Kaladhar Voruganti
Technical Director
Advanced Technology Group
NetApp, Inc.
Kaladhar.Voruganti@netapp.com

John D. Strunk
Member of Technical Staff
Advanced Technology Group
NetApp, Inc.
John.Strunk@netapp.com

ABSTRACT
Over the last 20 years, there have been many changes in the data storage industry. NetApp® products have kept pace and pushed the boundary in various areas. Staying at the forefront requires attentiveness to emerging technology trends and a disciplined approach to analyzing them. By understanding the trends and how they affect our customers, we can focus our efforts on delivering the best products possible. In this issue of OSR, we highlight some of the research and innovation that have helped us stay at the forefront of these technological changes.

1. INTRODUCTION
We recently celebrated the 20th anniversary since NetApp (www.netapp.com) was founded in 1992. This also marks two decades since the initial release of the Data ONTAP® operating system. Today, Data ONTAP is the #1 storage operating system in the world, running on several hundred thousand systems world-wide handling a wide variety of workloads and data types. The initial releases of Data ONTAP were focused on supporting file sharing for workgroups using the NFS protocol. Data ONTAP today supports all major storage protocols and runs on a wide variety of NetApp FAS systems (and IBM branded N-Series) and solution configurations.

Variations of the Data ONTAP operating system have also been used at various times for derivative products from NetApp including NetCache® web acceleration appliances, FlexCache® storage appliances, V-Series open storage controllers (an early form of storage virtualization), Data ONTAP Edge (a version of Data ONTAP that runs on a hypervisor) and StoreVault® storage systems for the MSE market.

Many papers have been published explaining various innovations in Data ONTAP through the years, covering topics such as:

- The seminal initial paper on WAFL® (Write Anywhere File System) [19].
- Multi-protocol support, initially adding support for CIFS as well as NFS [22], then adding iSCSI and FCP block protocols.
- The addition of clustering to Data ONTAP [25].
- Storage efficiency techniques such as block de-duplication, thin provisioning [17], and volume cloning.
- Reliability techniques such as dual parity RAID (RAID-DP® [23]), write-loss protection [24], and multi-path I/O for high availability.
- Replication for mirroring, disk-to-disk backup techniques and disaster recovery [20, 21].
- Caching, including recent advances in use of flash technology both in the storage controller and in the application server to accelerate performance [8, 9].

2. TECHNOLOGY TRENDS
Technology trends continually shape and drive the evolution of the IT industry. To stay at the forefront of these changes, we engage in research and prototyping efforts in many areas. NetApp invests significant resources in understanding the major trends in three ways.

Understanding the technology behind the trend
We conduct structured experiments so we can clearly measure the effects and capabilities; learn about the emerging ecosystem surrounding the technology; and investigate possible deployment alternatives.

Exploring the implications for our customers
It is important to recognize that changes in technology can enable powerful new solutions for customers, but it can also introduce new problems not previously faced. For example, the emergence of powerful mobile devices has had a significant effect on productivity and collaboration, but it has created new problems for security and geo-scale data delivery.

Examining how the new technology can enhance our products
This requires that we explore both the potential opportunities and the potential threats that will arise in our product portfolio, our competitor's product portfolios, and newly emerging competitors.

This deep understanding of the technology trends and their implications for our customers is used to guide future investments. Not all innovation presents as product capabilities, of course. NetApp also shares technical advances in the form of contributions to open source initiatives, participation in industry associations and standards groups, and publishing papers to help advance the state of the art in systems research. For example, the storage industry is experiencing key inflection points due to the emergence of many disruptive trends such as server virtualization, the emergence of flash, scale-out cloud storage architectures, and the use of data analytics to better manage systems. In order to help

the storage community better understand these trends, NetApp has made contributions, both through our own investigations as well as through collaborations with other researchers. Three main areas in which we have been involved are:

The evolving storage ecosystem
In order to design new storage systems, it is necessary for a storage architect to have a good understanding of new application workloads, emerging middleware and hypervisor architecture trends, and the changing characteristics of hardware resources like disks, CPUs, and networks. We have published three papers [1, 2, 3] that describe the characteristics of storage system workloads in real customer environments. We have also published a paper [4] about how to use database workload information to perform better data layout. During investigations of running Data ONTAP in a hypervisor, we explored efficient communication alternatives for inter-VM communication [5]. Finally, NetApp published a paper [6] that thoroughly analyzed the latent sector fault characteristics of disk drives by mining a large collection of support data.

Storage system innovations
Since the initial seminal paper on WAFL [19], increasingly storage systems are being asked to support new types of workloads. Thus, in addition to understanding the characteristics of these new workloads, it is important to make the appropriate architectural changes in the storage system I/O path. We published a paper [7] that examined the required architectural changes to support Hadoop workloads. Similarly, a lot of effort in the storage systems area is currently being devoted to finding new ways to leverage flash memory. In this regard, we have published two papers. The first discusses one way to leverage flash at the host server [8], and the second [9] discusses how to dynamically control the amount of flash given to a workload inside a storage controller. We have also published a paper that discusses how to use flash to increase the efficiency of disk I/O. NetApp has been a leader in the area of storage efficiency technologies for primary data, and we are continuing to innovate in this area. We published a paper on in-line deduplication for primary data, and another on selective video encoding [12]. We also published a paper on how to do thin provisioning in a performance efficient manner [17] and tracking back references in WAFL [18].

Storage management
It is becoming increasingly difficult to manage systems at scale. In order to address this problem, we have proposed a framework [13] on how to manage storage systems by using technology independent service level objectives (SLOs). NetApp is also doing research in how to mine system logs to better manage storage systems. We published multiple papers that analyze system logs to perform system fault diagnosis [14, 15, 16].

3. PAPERS PRESENTED

In this edition of OSR, we have chosen to present a small selection of papers that highlight less well known areas of innovation, not all of which have been released commercially.

The papers presented here are a combination of both some older work that we have not previously presented publicly, as well as new work. The earlier papers: RAID triple parity (2006), Glitz: Cross-vendor federated file systems (2008), Designing a fast file system crawler with incremental differencing (2008), and Hybrid Aggregates: Combining SSDs and HDDs in a single storage pool (2009) were all previously published in our NetApp Technical Journal for an internal audience. We have chosen to make them publically available, along with the remainder of the papers which are new work from recent projects.

The first group of papers in this issue is broadly related to storage management. They describe work on topics of cross-vendor interoperability, metadata indexing, performance modeling, and automatically responding to workload changes.

Glitz: Cross-vendor federated file systems
In version 4, the NFS protocol added the ability for a server to redirect clients during a pathname lookup. This was initially intended to support replication and migration, but it provided the opportunity to create a multi-vendor federated file system. The Glitz project describes NetApp's investigation into supporting such cross-vendor file systems.

Designing a fast file system crawler with incremental differencing
This paper presents the design of a utility to traverse file system metadata for use in indexing and search. It describes one design based on standard file system APIs and another that uses properties of the WAFL file system to speed the initial indexing as well as later incremental updates.

Model building for dynamic multi-tenant provider environments
Having accurate models of storage system performance is important for multi-tenant environments. Service providers need to share resources across tenants to lower system costs, but in order to efficiently share resources, it is necessary to be able to predict the level of service that will be provided. This paper discusses a machine learning based black box modeling algorithm that is designed to provide predictions of system performance.

Responding rapidly to service level violations using virtual appliances
The properties of storage workloads (e.g., working set sizes and request rates) change over time, and these changes can lead to SLO violations or significant resource over-provisioning. This paper describes an investigation into dynamically instantiating virtual storage appliances to handle spikes in the storage workload. The Dynamite system monitors I/O performance and creates virtual caching appliances, as needed, to service the load.

The second group of papers deals with the area of storage efficiency. They describe work on enhancing parity protection, dynamically combining SSDs and HDDs, and using variable length deduplication.

RAID triple parity
The original Row Diagonal Parity paper was published in FAST 2004. It described a double parity scheme, allowing RAID groups to survive up to two disk failures (or more commonly, one disk failure plus a latent sector error on another disk). The paper presented here extends the RDP algorithm to allow triple parity, allowing a RAID group to survive up to three disk failures.

Hybrid Aggregates: Combining SSDs and HDDs in a single storage pool
The Hybrid Aggregates project was one of a number of projects that examined ways of integrating flash-based storage into NetApp's products. It looked at combining SSDs and HDDs in a single pool of storage, with data moving between the two classes of storage transparently to the user. This paper describes early work that laid the foundation for NetApp's Flash Pools.

Space savings and design considerations in variable length deduplication
This paper investigates the potential savings of using hierarchical data deduplication for object-based workloads. Hierarchical data deduplication allows data to be deduplicated at widely varying granularities, leading to potentially significant space savings over standard data chunking techniques.

3. LOOKING FORWARD

Technology trends are driving product evolution faster than ever. We are in the midst of major transformations driven by solid state storage such as flash, virtualization of servers and networks, ubiquitous high-speed, low-cost networks (LAN, WAN, wireless), and the emergence of a multitude of applications and services delivered from cloud providers. Data types are evolving in new forms such as streaming video and machine generated analytical information with very different characteristics and data management requirements. Mobile devices are driving requirements for data and applications to be accessible "anywhere, anytime." Finally, new styles of application architectures are changing the methods of accessing data and the very fundamental notions of what "I/O" is.

NetApp will continue to innovate and continue to deliver solutions to the challenges our customers face in managing and harnessing their vast and rapidly growing mountains of data.

4. REFERENCES

[1] Chen, Y., Srinivasan, K., Goodson, G., and Katz, R. 2011. Design Implications for Enterprise Storage Systems via Multi-Dimensional Trace Analysis. In SOSP.

[2] Yadwadkar, N.,Bhattacharyya, C., Gopinath, K., Thirumale, N., and Susarla, S. 2010. Discovery of Application Workloads from Network File Traces. In USENIX FAST.

[3] Leung, A., Pasupathy, S., Goodson, G., and Miller, E. 2008. Measurement and Analysis of Large Scale Network File System Workloads. In USENIX FAST.

[4] Oguzhan, O., Salem, K., Schindler, J., and Daniel, S., 2010. Workload Aware Layout for Storage Systems, In ACM SIGMOD.

[5] Burtsev, A., Srinivasan, K., Radhakrishnan, P., Bairavasundaram, L., Voruganti, K., Goodson, G. 2009. Fido: Fast Inter-Virtual-Machine Communication for Enterprise Appliances. In USENIX ATC.

[6] Bairavasundaram, L., Goodson, G., Pasupathy, S., and Schindler, J. 2007. An Analysis of Latent Sector Errors in Disk Drives. In ACM SIGMETRICS.

[7] Mihailescu, M., Soundararajan, G., and Amza C. 2012. MixApart: Decoupled Analytics for Shared Storage Systems. In USENIX HotStorage.

[8] Byan, S., Lentini, J., Madan, A., Pabon, L., Condict, M., Kimmel, J. Kleiman, S., Small, C., and Storer, M. 2012. Mercury: Host Side Flash Caching for the Data Center. In IEEE MSST.

[9] Sehgal, P., Voruganti, K., and Sundaram, R. 2012. SLO-Aware Hybrid Store. In IEEE MSST.

[10] Schindler, J., Shete, S., and Smith, K. 2011. Improving Throughput for Small Disk Requests with Proximal I/O. In USENIX FAST.

[11] Srinivasan, K., Bisson, T., Goodson, G., and Voruganti. K. 2012. iDedup: Latency-aware, Inline Deduplication for Primary Storage. In USENIX FAST.

[12] Khatpal, A., Kulkarni, M., and Bakre, A., 2012. Analyzing Compute versus Storage Tradeoff for Video-Aware Storage Efficiency. In USENIX HotStorage.

[13] Bairavasundaram, L., Soundararajan, G., Mathur, V., Voruganti, K., and Kleiman, S. 2011. Italian for Beginners: The Next Steps for SLO-Based Management. In USENIX HotStorage.

[14] Yin, Z., Ma, X., Zheng, J., Zhou, YY., Bairavasundaram, L., and Pasupathy, S. 2011. An Empirical Study on Configuration Errors on Open Source Systems. In SOSP.

[15] Yuan, D., Mai, H., Xiong, W., Tan, L. Zhou, YY., and Pasupathy, S. 2010. SherLog: Error Diagnosis by Connecting Clues from Run-Time Logs. In ASPLOS.

[16] Jiang, W., Hu, C., Pasupathy, S., Kanevsky, A., Li, Z., and Zhou, YY. 2009. Understanding Customer Problem Troubleshooting from Storage System Logs. In USENIX FAST.

[17] Edwards, J., Ellard, D., Everhart, C., Fair, R., Hamilton, E., Kahn, A., Kanevsky, A., Lentini, J., Prakash, A., Smith, K., and Zayas, E. 2008. FlexVol: Flexible, Efficient File Volume Virtualization in WAFL. In USENIX ATC.

[18] Macko, P., Seltzer, M., Smith, K.2010. Tracing Back References in WAFL. In USENIX FAST.

[19] Hitz, D., Lau, J., and Malcolm, M. 1994. File System Design for an NFS File Server Appliance. In USENIX Winter Conference.

[20] Patterson, H., Manley, S., Federwisch, M., Hitz, D. Kleiman, S., and Owara. S. 2002. SnapMirror: File System Based Asynchronous Mirroring for Disaster Recovery. In USENIX FAST.

[21] Hutchinson, N., Manley, S., Federwisch, M., Harris, G., Hitz, D., Kleiman, S., O'Malley, S. 1999. Logical versus Physical File System Backup. In USENIX OSDI.

[22] Pawlowski, B., Juszczak, C., Staubach, P., Smith, C., Lebel, D., Hitz, D. 1994. NFS Version 3: Design and Implementation. In USENIX Summer Conference.

[23] Corbett, P., English, R., Goel, A., Grcanac, T., Kleiman, S., and Sankar, S. 2004. Row-Diagonal Parity for Double-Disk Failure. In USENIX FAST.

[24] Krioukov, A., Bairavasundaram, L., Goodson, G., Srinivasan, K., Thelen, R., Arpaci-Dusseau, R., Arpaci-Dusseau. A. 2008. Parity Lost and Parity Regained. In USENIX FAST.

[25] Eisler, M., Corbett, P., Kazar, M., Nydick, D., and C. Wagner. 2007. Data ONTAP GX: A Scalable Storage Cluster. In USENIX FAST.