:login; juin 2ØØ7, WiP report du FasT'Ø7

■ *CompulsiveFS: Making NVRAM Suitable for Extremely Reliable Storage*

Kevin Green presented CompulsiveFS, a file system that stores persistent metadata directly to an erasure coded log in NVRAM instead of a RAM cache. The file system performs incremental erasure-encoding and signature computations over the contents of the log to create fault tolerance. The current prototype log is an order of magnitude faster than page-protected caches for small writes of 10–50 bytes. CompulsiveFS is in the early stages of research, design, and implementation.

■ *Performance Evaluation of RAID6 Systems*

Yan Li presented a plan for a three-piece study on the performance characteristics of RAID6 under the Storage Performance Council-1 benchmark. The study will simulate the storage using SimRAID, which models and simulates the RAID controller, cache, fibre-channel bus, and disks and has been shown to have a maximum inaccuracy of 5%. The study looks at RAID6 performance under fault-free mode, degraded mode, and recovery mode. Of particular interest is finding the ideal mix of handling requests versus rebuilding a disk.

■ *GANESHA, a Multi-Usage with Large Cache NFSv4 Server*

Philippe Deniel presented work on GANESHA, a user-space NFSv4 server with a large cache and support for a number of backend file systems. GANESHA has a filesystem abstraction layer (FSAL) that makes writing plug-ins for different backing file systems easy. Currently, there is support for HPSS and POSIX backends. A number of interesting backends are under development; these include an NFSv4 client that will allow GANESHA to be a proxy, an SNMP backend that will allow MIBS to be browsed, and the LDAP backend that will allow browsing of trees. Announcements to SourceForge and freshmeat are forthcoming.

■ *FlexiCache: A Flexible Interface for Customizing Linux File System Buffer Cache Replacement Policies*

Pavan Konanki presented the initial work for FlexiCache, an interface in the Linux kernel to accommodate different buffer cache replacement algorithms. The motivation for this work is to test new replacement algorithms, such as ARC, PCC and LIRS, that have been shown to perform better under certain access patterns. Also, this API would accelerate the testing and development of new buffer cache replacement algorithms. The key issue is trying to design the system to support many replacement policies while keeping the cache mechanics hidden. The performance implications of this added generality are also unknown.

■ *Diamonds Are Forever, Files Are Not*

Surendar Chandra talked about storage systems that use an "importance number" to decide how to manage a data store automatically. The experiments were motivated by a storage server for video-recorded lectures where some of the data may be less important than new data coming in and can be removed. To make the system successful, administrators must be able to specify accurate object lifetimes; therefore providing usage feedback is important. Currently a system is being built to prototype this idea.

■ *RBF: A New Storage Structure for Space-Efficient Queries for Multidimensional Metadata in OSS*

Yu Hua presented RBF, an r-tree with a bloom filter at each node, a structure that allows for efficient point and range queries. Point queries ask whether an object is in a data set and a range query grabs the set of objects that match a query. The application of this is to make efficient object-based storage devices that can do point/range operations on object metadata. Currently, a real 10TB storage system has been implemented using a partial implementation of the RBF structure.

■ *Storage Performance Isolation: An Investigation of Contemporary I/O Schedulers*

Sarala Arunagiri presented research on the performance isolation characteristics of a number of modern I/O schedulers. This research is motivated by the quality of service guarantees that large consolidated shared storage must make. The findings suggest that many I/O schedulers do not provide performance isolation in all circumstances.

INVITED TALK

■ *Trends in Managing Data at the Petabyte Scale*

Steve Kleiman, Senior VP and CTO, Network Appliance

Summarized by Michael Mesnier
(michael.mesnier@intel.com)

In the first three-quarters of 2006, approximately 900 PB of storage was shipped worldwide by storage systems vendors, including Dell, EMC, Hitachi, HP, IBM, and Network Appliance. Of this total, Network Appliance shipped approximately 200 PB and is currently at a 100 PB/quarter run rate.

The big challenge today is keeping pace with such growth. A 50–100% yearly increase in storage capacity is not uncommon. Most of this growth is from unstructured data (e.g., contracts, letters, memos) and semi-structured data (e.g., email), but structured data (db) is also growing. Managing this growth introduces hidden burdens (costs).

A common strategy is to overprovision, resulting in low disk utilizations (with 25% or less being typical). Of course, CIOs do not want to take chances when it comes to safely storing company and/or customer data. Today's legal burdens (e.g., regulatory compliance) and social burdens (e.g., losing data is bad press) underscore this point. In general, overprovisioning stems from the "build-out" manner in which most storage infrastructures are managed today (i.e., within application-centric "silos"), including