CSE726
Data Intensive Distributed Computing
Spring 2011

A Review of Panache: A Parallel File System for Global File Access
M. Eshel et al,.FAST '10

Presented by: Ajinkya
Reviewed by: Rishi Baldawa

Panache is an IBM product which is scalable, high performance and has clustered file system cache; aiming to provide flawless access to remote datasets using POSIX interface. It can be used for parallel data intensive applications that require Wide Area Network. In this file system, every aspect of the architecture is parallel in nature such as ingest, access, updates and write-backs. It takes care of the Wide Area Network latencies and outages using asynchronous operations employ conflict handling and resolution in disconnected mode operations.

Panache uses pNFS along with GPFS. GPFS is used as the high performance storage cluster system while pNFS protocol reduce bottlenecks by using GPFS storage protocols and provide direct access to GPFS. "Panache is implemented as a multi-mode caching layer, integrated within the GPFS that can persistently and consistently cache data and metadata from remote cluster." Within a cache cluster, every node can access every cached data and metadata providing applications (running on the same cluster) with the same performance as that on server where the data and metadata are actually located. Panache allows asynchronous updates of the cache for improved application performance on the local machine. The Cache cluster architecture has two types of files systems: Cache Clusters and Remote Cluster File Systems. The nodes on these systems can be of two types as well: Application nodes which service application data requests and Gateway Nodes which act as the client proxies to fetch data in parallel from remote site and store it in the cache.

Panache uses Local Consistency, Validation Lag, Synchronization Lag and Eventual Consistency to ensure (and fine tune) consistent data at local and remote site. Panache applies distributed locking mechanism to make sure that the cache cluster is always locally consistent for updates. All data and metadata read operations are synchronous to ensure the data or metadata is cached and is valid. Panache ingests the data and metadata in parallel from multiple gateway nodes so the speed is only limited by network bandwidth. Asynchronous operations are used for all data and metadata updates. It was observed that if three commands are pair-wise ordered, then the three commands form a time ordered sequence and that if Objects are pair-wise dependent, then even the first and the last object are dependent. Inode scans are used for recovery purposes. Access Control Locks setup at the remote cluster are enforced at the cache. Panache also performs conflict handling.

The performance is measured based on I/O Performance (to check storage system throughput), Metadata Performance (to measure metadata update performance) and WAN Performance (to validate the effectiveness of the Panache over WAN). The overall performance of Panache is very good in comparison however it system doesn't come without shortcomings most of which were discussed during the presentation.

During the presentation, following points were discussed. It is important to notice that WAN is supported because WAN latencies are covered. The definition of Local Cluster or Local System as it can mean either a system without cache, or one inclusive of cache. Inode on the remote side can clash with the Inode on the client side and hence such conflicts need to be handled. Even asynchronous operations need conflict handling. However, the paper fails to describe in detail, how it performs conflict handling in general. There is also no prefect policy and this operation has to be done manually. The system doesn't prevent conflict but can only solve them, once they prop up. There was not much discussion on the overall performance of the system due to time constraints.