

DMF for Lustre Update

Mark Seamans
mseamans@sgi.com
February 2016

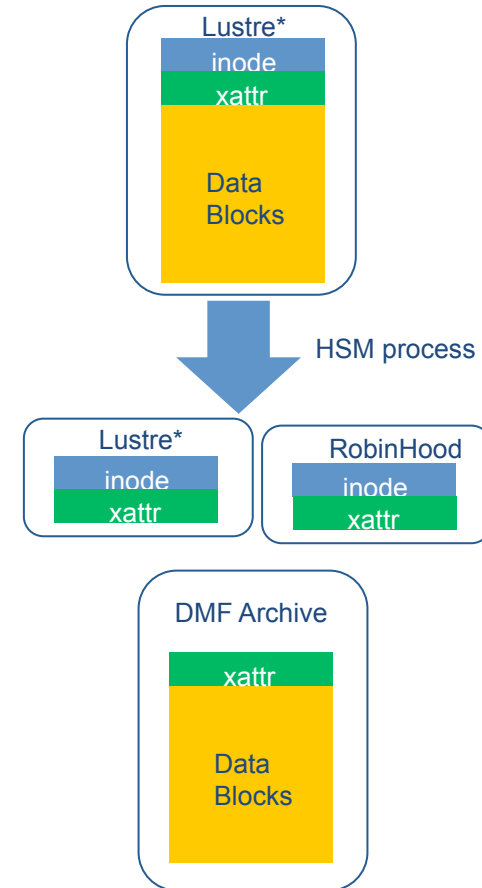


Agenda

- DMF for Lustre Summary Overview
- Data Flow and Architecture
- Status and Best Practices
- RobinHood Status and Roadmap
- DMF Policy Engine Roadmap
- Discussion

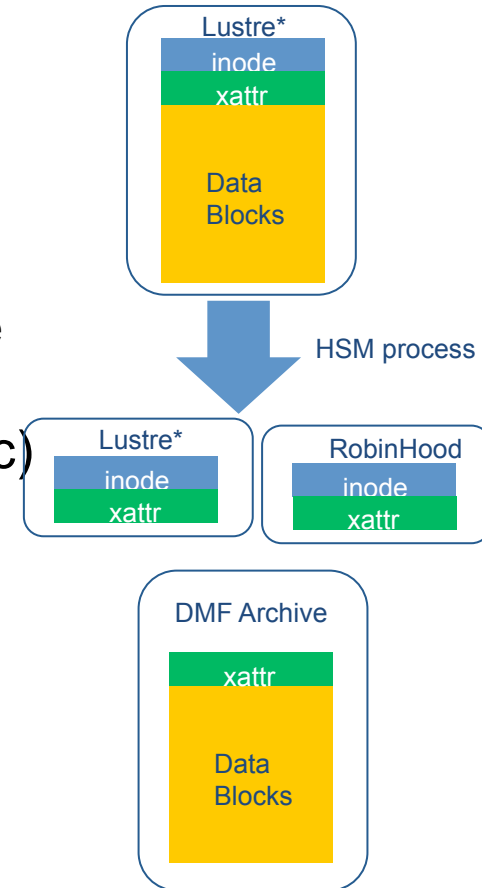
DMF for Lustre Summary

- HSM support included as a feature of Lustre since Lustre 2.5
- Three basic components to a Lustre* file:
 - inode (metadata – permissions, times)
 - xattr (extended attributes – striping layout (lov))
 - Data blocks (data)
- Lustre* DMF HSM archives the data blocks and the xattr information
 - Inodes are asynchronously updated in Robinhood Policy Agent



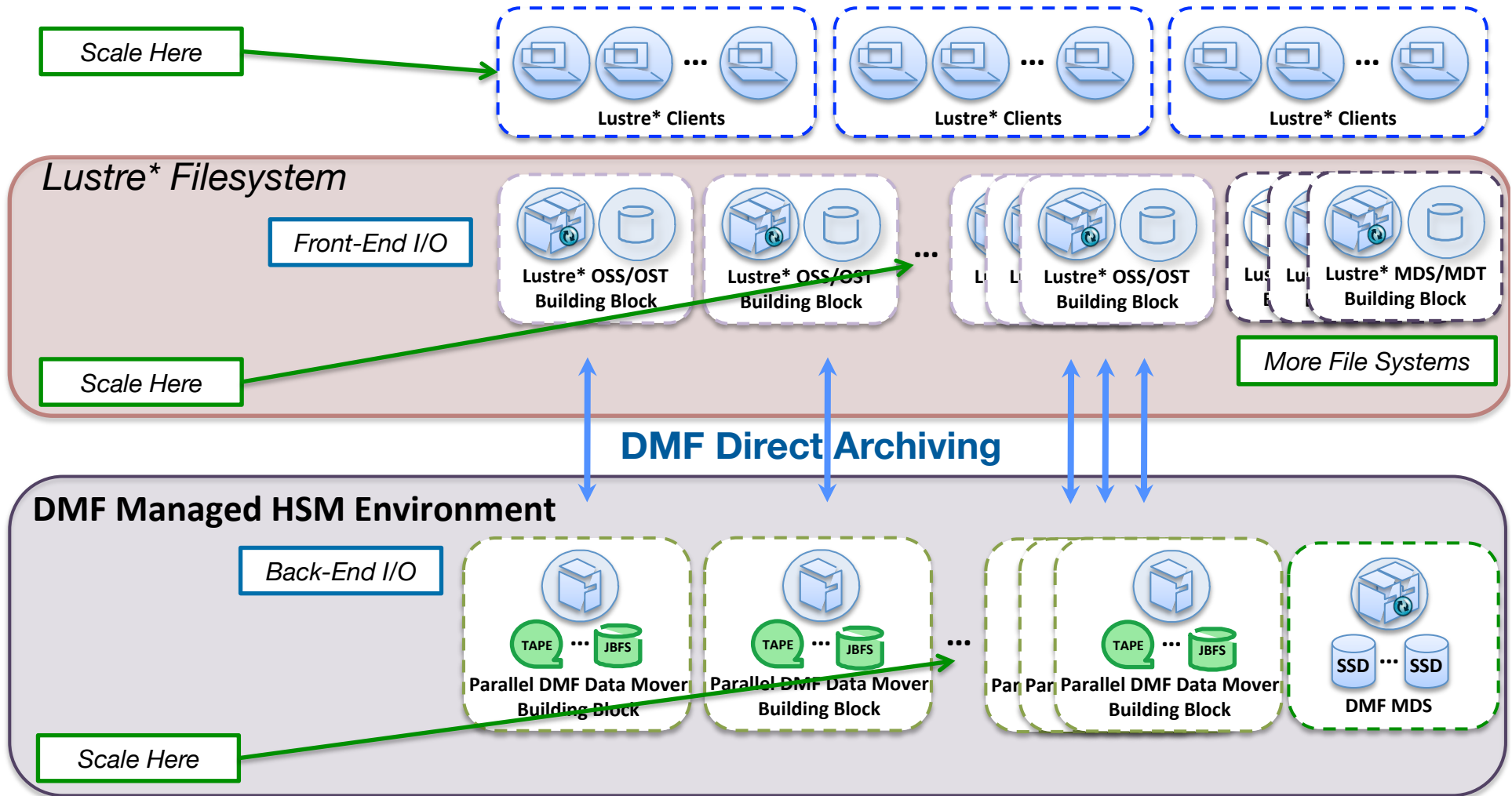
DMF for Lustre Summary

- A file stub and the xattr information stays in place on Lustre*
 - inode stays on primary storage (about 2K in size)
 - Typically this is called a stub
 - After a period of time, the data blocks on primary storage are released
- Files can be “restored” when needed (automatic)
- Lustre* has a fixed stub size (inode size ~2K)
- The Robninhod Policy Agent keeps a copy of the Inode
 - Asynchronous updated
 - Adds data protection and a recovery mechanism for Lustre
 - RobinHood is included as part of Intel EE for Lustre and supported as part of EE for Lustre support agreements



DMF for Lustre

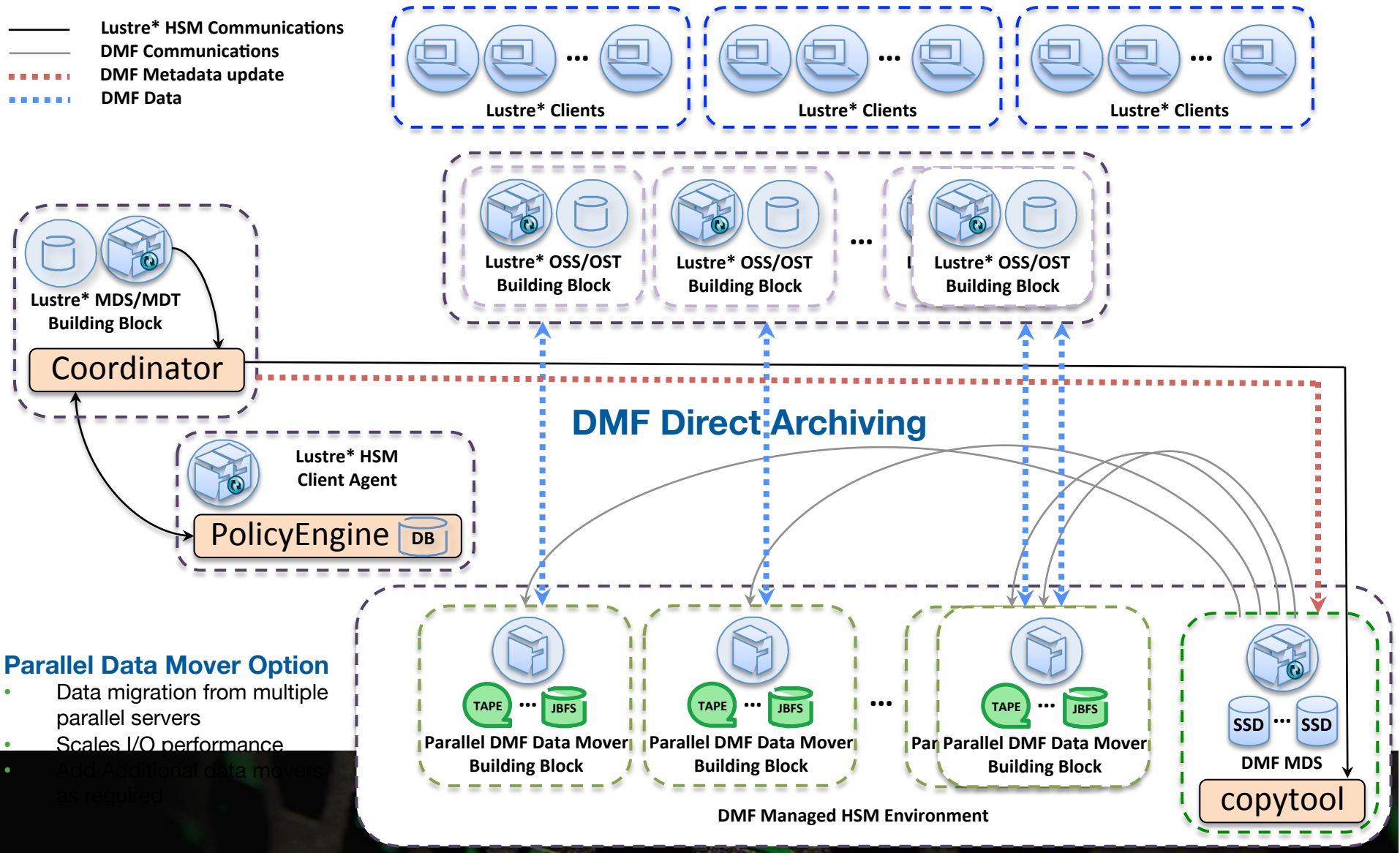
Capacity, Performance & Reliability



Lustre* HSM | Communication & Data Flow

* = Some names and brands may be claimed as the property of others

- Lustre* HSM Communications
- DMF Communications
- ⋯ DMF Metadata update
- ⋯ DMF Data

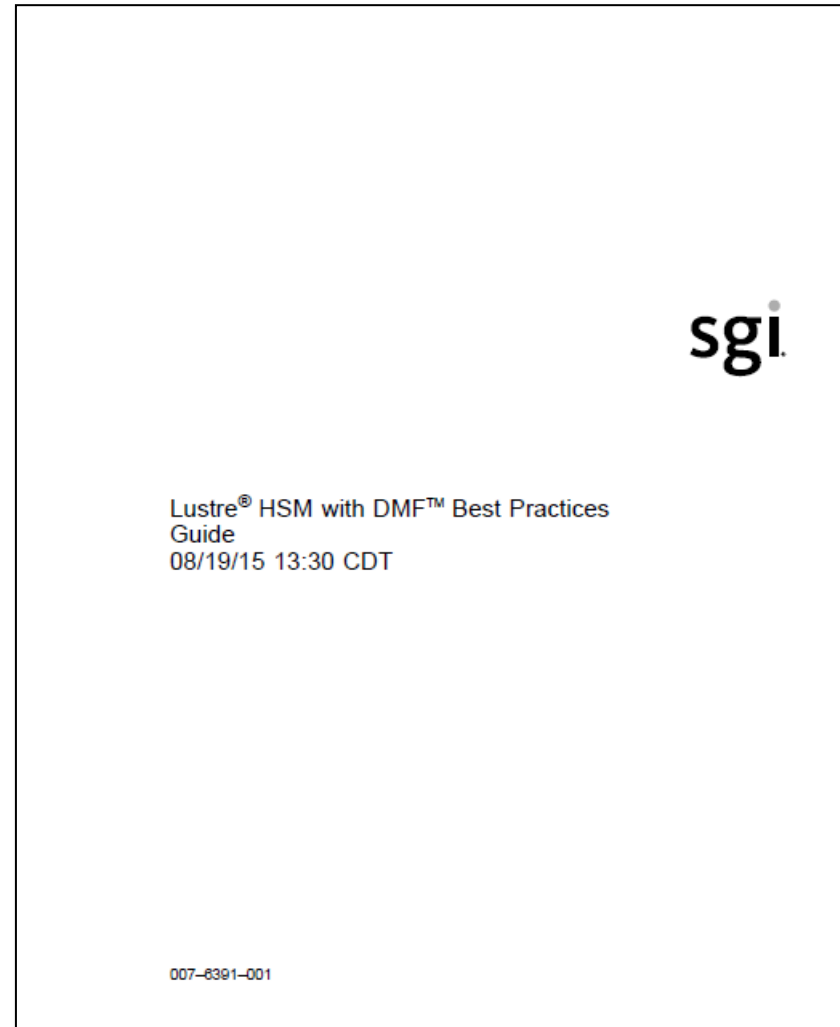


Parallel Data Mover Option

- Data migration from multiple parallel servers
- Scales I/O performance
- Additional data mover required

Status and Best Practices

- Several clients sites in production
- RobinHood policy engine has required focus & attention
- SGI Best Practice Guide available
 - Install, tuning, RH file recovery commands and process



Status and Best Practices

- SGI Storage Solutions team performing extended validation and best practice development now
- Updated guidelines for RH server configuration (and updated RH software) made available in Q4 2015
 - Focus on RH server memory and storage I/O performance
 - Current comfort zone is for file systems up to the ‘hundreds of millions’ of files object count
 - Updated SGI Best Practice Guide coming (contact SGI if planning a deployment)

RobinHood Status & Roadmap

- Many updates to RobinHood and Lustre HSM code over the last year
- RobinHood v3.0 Alpha released in December 2015
 - Currently in test
 - Most/all performance related features have been back-ported to RobinHood 2.5.5
- Database and scalability items planned for 3.X release

RobinHood Best Practices

Hardware Setup

- 4x Object Storage Servers
 - 2x Intel® Xeon® E5-2680 v4
 - 64GB of RAM.
 - 4x Intel® SSD DC P3701
 - Mellanox® FDR card
- 1x Metadata Server:
 - 2x Intel® Xeon® E5-2680 v4
 - 64GB of RAM.
 - 2x Intel® SSD DC P3701
 - Mellanox® FDR card
- 1x Policy Server:
 - 2x Intel® Xeon® E5-2680 v4
 - 64GB of RAM.
 - 2x Intel® SSD DC P3701
 - Mellanox® FDR card

How to design the RBH server


- High frequency metadata operations
- As much metadata as possible
- 80% of available innoDB buffer pool

First scan operation

First scan operation with RBH on the same server. In the chart metadata operations per second collected by Intel Manager for Lustre*

DMF for Lustre Evolution

- Lustre HSM integration will continue to leverage RobinHood in the near term
- An upcoming release of DMF (details in DMF Roadmap session) will incorporate native capabilities for log processing, HSM and file system recovery tools



SGI DMF™ for Lustre Data Sheet

SGI InfiniteStorage DMF for Lustre
Extending Lustre Capacity With Policy-based Tiered Data Management

Key Features
Policy Management & Active Data Migration
Storage Virtualization & Tiered Data Management
Cold Storage Integration with SGI ZeroWatt™ Power Control

Expanding the Capacity and Value of Lustre
Building on SGI's multi-decade strength as a leader in tiered data management solutions, SGI DMF for Lustre allows organizations to expand their use of Lustre file systems to include long term data storage in addition to being the file system of choice for high performance data access within cluster compute environments.
With DMF for Lustre, active data is always available on the fastest tier of Lustre storage (OST) while tunable policies allow the SGI solution to automatically migrate data to and from additional tiers of storage that can include lower cost disk, public and private cloud and even robotic tape libraries. The result is a system where all data is online and available for access at all times - but where storage costs are managed and minimized based on the organization's objectives and data usage patterns.

Broad Support for Storage Technologies
SGI's DMF supports storage tiers that can include Storage Area Network (SAN) arrays and Network Attached Storage (NAS) devices from a broad range of open systems suppliers. Object-based storage and cloud-based storage that leverage Amazon S3 interfaces are also supported in both private cloud and public cloud models using adapters that are created, validated and supported by SGI.
Additionally, SGI is pioneering the delivery of integrated hardware-and-software capabilities for intelligently managed cold storage with extremely low storage cost characteristics while also delivering operational savings through software-managed ZeroWatt™ power control that can significantly reduce the power and cooling costs for disks containing less active data. And for more static "dark data", DMF supports a wide array of library-based tape storage platforms that can deliver very appealing cost per terabyte levels while maintaining the data in a ready-for-access state.

Diagram: A layered architecture diagram showing Lustre HPC Clients (Lustre HPC Client, Lustre SGI I/O Client, Lustre Hadoop Client) connected to a Lustre Storage Environment (MDS / MDT, OSS / OST). This environment is integrated with SGI DMF™ (Lustre HSM and Data Management), which in turn connects to various Storage options (Disk - Tape - Cloud - SSD - Cold Storage). Arrows labeled 'EXTEND' indicate the integration points.

SGI DMF for Lustre: Direct Integration with Lustre Storage Environments

sgi.

DMF for Lustre Integration

- Available today as part of DMF 6
- Will be reintroduced in v7.X
- DMFv7 will process Change Log directly
 - No requirement for RobinHood
- SGI will implement Change Log and HSM coordinator API as a standard
 - Also apply it to CXFS / XFS integration
 - Roadmap session will also discuss thoughts around GPFS

Summary

- Work with SGI for any planned DMF for Lustre deployments
- Solution has gone through significant testing and tempering
- Near term documentation and RobinHood updates coming
- Longer term SGI native policy engine support will drive scalability and tighter integration

sgi®