STORAGE DEVELOPER CONFERENCE



Virtual Conference September 28-29, 2021

Fully Autonomous Storage and Memory Hierarchies

The Future of Storage

Presented by Irfan Ahmad, CEO and Founder, Magnition.io



Magnition.io (Founder, CEO)
CloudPhysics (Founder, CTO, acq. by HPE)
VMware (Kernel, Storage DRS, IPO)
Transmeta (x86 Processor, IPO)
40+ Patents, 11 peer-reviewed papers, 3 best paper awards
University of Waterloo, Math/CS
@virtualirfan





Does the Industry Need Fully Autonomous Storage?





This is you

This is your customer







Highly Dynamic Domains Benefit from Full Autonomy













Manual Storage / Memory Management Now Infeasible

Applications and data requirements changing hourly.



Manually-managed Storage / Memory Infrastructure Vulnerable to:

- Thrashing, Scan pollution
- Gross unfairness, Interference
- Unpredictable availability
- Data loss risks

- \Rightarrow Overprovisioning
- \Rightarrow Lack of Control
- ⇒ Availability & Durability Risk



Time has Come for Fully Autonomous Storage and Memory Hierarchies



Autonomous Storage ML/Models Needed





Self-Awareness

Caches, memories, disks, data paths, latencies, link throughput limitations, media costs, data movement costs, performance capabilities, degraded performance, etc.

Environment Awareness

Dynamic workloads, QoS constraints, competing traffic on links, dynamic laaS costs, failures, imminent failures, flash wear/tear, power constraints, temperature, dynamic resource costs, etc.

Self-Awareness

Acceleration, braking steering, roll, wear/tear, weight distribution, battery discharge temperature and load models

Environment Awareness

Maps, static obstacles, dynamic obstacles, object capabilities, terrain, distances, relative object velocities, live traffic, GPS, road conditions, weather, law enforcement, etc.



Autonomous Systems Require OODA Loops & Models









Architecture for Fully Autonomous Storage





Fully Autonomous Storage / Memories are Self-Aware



Fully Autonomous Storage Needs Must Continuously Answer

• Is this performance good?



- Can performance be improved?
- How much Cache for App A vs B vs ...?
- What happens if I add / remove DRAM?
- How much DRAM versus Flash?
- How to achieve 99%ile latency of X μs?
- What if I add / remove workloads?
- Is there cache thrashing / pollution?
- What if I change cache parameters?



Use Case #1: Autonomous QoS SLA

- How?
 - Users dial-in latency or throughput target and budgets
 - Fully Autonomous Storage auto allocates just enough capacity to meet SLAs at all times
- Value for Customer
 - Automated SLA achievement!
 - Set and Forget, ease of mind
 - Revenue disruption avoidance
 - Improved margins
 - Zero OpEx performance scaling
 - Dramatically reduced service interruptions





Use Case #2: Autonomous Cost / Performance Optimization

- How?
 - Real-time workload modeling
 - Resource allocation predictions
 - Dynamic resource adjustment and isolation
 - Auto right-sizing
- Value for Customer
 - Lowest total cost of ownership (TCO)
 - Eliminate noisy neighbor problems
 - Policy-driven operations
 - Lower OpEx for infra teams
 - Predictive planning

Cache Size & Latency Reduction (Thrashing Remediation)







Technical Breakthroughs In Performance Autonomy for Fully Autonomous Storage and **Memory Hierarchies**



Fully Autonomous Latency Targets





Fully Autonomous Latency Targets





Fully Autonomous Multi-Tier Allocation



* Can model network bandwidth as a function of cache misses from each tier



Fully Autonomous Storage is Within Reach





This is you

This should be your customer



17 | ©2021 Storage Networking Industry Association. ©2021 Magnition, Inc. All Rights Reserved.

https://unsplash.com/photos/9xMf-zZu6kM



irfan@magnition.io @virtualirfan



18 | ©2021 Storage Developer Conference ©. Insert Company Name Here. All Rights Reserved.