### Unlock the Data Center





# **Disrupting Key-Value**

**Abstract:** 

Key-value storage underpins all large-scale web applications, consuming hundreds of thousands of servers with substantial power and cost implications. As data center trends shift from traditional compute and software-defined solutions to accelerated compute, key workloads like AI and networking are adopting hardware acceleration, but storage lags behind. This talk will address current data center challenges and argue that hardwareaccelerated key-value storage offers a superior path compared to computational storage, promising enhanced efficiency and performance.

#### Speaker: Andy Tomlin: CEO, Chief Architect, Founder



Executive at Sandisk, Sandforce, WD, Samsung, Kioxia, 30yrs Storage experience. ~60 patents, many Storage, Flash and SSD products delivered to Market. Start up experienced (Sandforce exec, Agoraic founder) https://www.linkedin.com/in/andytomlin/



Storage Developers Conference 2024

### **History of Storage Devices**

- Since the beginning of Storage, devices have used fixed-capacity blocks
- For over 50 years this has been 512 byte block size (+ some variants)
  However, the things we store are not fixed size, and are either smaller than a block or larger than a block.



## Mapping systems

iStor



### **Optimal solution**

**Axiomatic best solution** 

Perform allocation of space and location tracking in one place

Lowest Level

Best Performance

Perform mapping in

Hardware

Does not mean eliminating higher level functionality above, just that those systems deal with abstracted locations that don't need to change Single garbage collection Single overprovisioning Variable sized data Variable number of items



## **Implication of Optimal Solution**



 The capability of today's HW means that this problem can be solved in HW if we have the will and creativity to do it



### **A Different Storage Abstraction**

What is Key-Value?

iStor

An abstract user defined reference to variable length data

Key Value

Why is this useful? An application can generate the key from other reference data without needing to tracking it





Mona Lisa

### **Key-Value Storage Underpins the Web**



### Data Centers Must Become More Power Efficient

The Washington Post Democracy Dies in Darkness

AI is exhausting the power grid. Tech firms are seeking a miracle solution.

The New Hork Times

BUSINESS

A.I. Frenzy Complicates Efforts to Keep Power-Hungry Data Sites Green

#### BARRON'S

Is AI A Major Drain On The World's Energy Supply?

The problem of Data Center power consumption is penetrating mainstream media. Data Centers will consume 8% of US power by 2030

#### B B C

Data centre power use 'to surge six-fold in 10 years'



Clean tech, AI boom straining US energy supply ChatGPT's boss claims nuclear fusion is the answer to Al's soaring energy needs. Not so fast, experts say

CINN





## Data Center Scaling is a fundamental Problem



iStor

### **AI Accelerates Demand for Key-Value Performance**

Training data, a large quantity of unstructured data consumed with high-performance

AI infrastructure, model storage Vector Databases running on top of Key-Value stores

Direct, Key-Value store to AI engine is an exciting possibility as a method to reduce CPU training bottlenecks The entire AI pipeline and management system solution heavily utilizes keyvalue Vector Database innovations need key-value storage engines to operate efficiently

AI demand is driving up power usage, putting enormous strain on existing Data Center Infrastructure, requiring new innovations for improved infrastructure efficiency



#### Data Center Workloads Shifting to Hardware Offloads



## Today's Software Solutions Create High Write-Amp

#### Log Structure Merge Tree



High write amplification due to WAL and SST level compaction High read amplification due to levels and file size L0 - 10MB, L1 - 100MB, L2 - 1GB ...

#### **Sample Config**

250TB System 500K read 100K write 1K value size average

#### **Sample Solution**

Nodes: 18 x i3en.24xlarge (96 vCPUs, 768 GB RAM, 58.95 TB storage) Disk storage: 1.03 PB Total vCPU: 1,728, Total RAM: 13,824 GB

iStor

# Criteria To Move Algorithms to Hardware

- Are the requirements well understood?
- Are they stable?
- Is there sufficient volume to enable commercialization?
- Is the problem important enough to solve?
- Will HW solution bring additional performance & power improvements?





## Path to an Optimal Key-Value Store



Use Less Flash & Better Performance

iStor

Green solution to scaling KV

**Customer Focused** 



# Key Value-as-a-Service

#### **The Problem**

Data Center Demand is growing faster than CPU scaling

#### **The Solution**

Disruptive HW-accelerated Key-Value Service



Database & Key-Value cost is driven by expensive compute

iStor

Utilize existing Cloud based FPGA infrastructure

Database Market size >\$100B WW, 15% CAGR

## How Does SW LSM compare to HW acceleration?

	LSM Vendor recommendations	HW accelerated
Storage	50% utilization (100% overprovisioned) (100TB = 200TB Flash)	75% utilization (33% overprovisioned (100TB = 133TB Flash)
DRAM	30 : 1 to Storage	N/A
Servers	1 vCPU per 12K operations	1 FPGA per 750K operations, max 8 FPGA per server

Significant Flash, Server, Memory, Power, Latency, Cost reductions



# New Offload Architecture has Breakthrough Benefit

<b>AEROSPIKE</b>				
	Par sarvar	Total ( <mark>20</mark>		
		servers)		
Capacity	60TB	1200TB		
Power <sup>1</sup>	2.91kW	58.2kW		
Write (kops)		753		
Read (kops)		3017		
Latency		1ms		
3yr TCO		\$4.17M		

Comparison of 1PB Solutions	
New	
Architecture	/

	iStor	
	Per server	Total (2 servers)
Capacity	Disaggregated	1000TB
Power <sup>1</sup>	2.91kW	5.8kW
Write (kops) <sup>2</sup>		1500
Read (kops)		6000
Latency		250uS
3yr TCO		~\$1M

https://aerospike.com/lp/y-running-operational-workloads/

#### **Key Takeaways**

**10x** - Less Hardware **10x** - Less Power

iStor

**10x** - TCO Reduction higher performance Better Reliability

#### CO<sub>2</sub> savings of 325 Metric Tons/yr

 Based on EPA power calculator, AWS outpost Data. Just rack power, does not include cooling power savings
 QiStor numbers based on model – 4 FPGA config + drive power estimate for Disaggregated

# **Solution Stack**

#### Key Goals

- Avoid proprietary interfaces
- Open-source Storage Engine
- Integrate with existing Databases and ORM's to minimize customer code changes
- NVMe KV interface via SPDK/DPDK



@gistor.com



#### **Customer Flow**



## Key Technical Challenges

- Developing complex Algorithms in HW (FPGA/ASIC)
- Typically takes large team 40-100 people including Architecture, Design, Verification, Firmware



Complexity

- Communication
- Keeping everything in synch, documentation
- Change management

### Arch2Code

iStor

- Describe architecture in YAML
  - Define all design elements: Interfaces, Memories, Registers, Functional Decomposition
  - Generate all Headers, Hierarchy, Implementation plumbing
- Create complete Reference SW implementation (SystemC)
- Port implementation from SystemC to SystemVerilog



# **Arch2Code - Testing**

- SystemC Reference Implementation is also the Test Harness
- Any part of the hierarchy can be replaced with RTL
  - Reference implementation provides test environment
- Any part of the hierarchy can be replaced with Paired RTL / Model self checking block
  - Significantly faster debug of RTL as expected response is provided by reference implementation



- QiStor team has proved methodology
  - Datapath width changes done in 1 day
  - RTL Design mapped to FPGA implementation in <1 week
  - AWS-F1 FPGA Demo with NVMe KV bring-up with no HW bugs
  - Debug of RTL much faster



### Summary

iStor

