KOVE™

MEMCON 2024

Software-Defined Memory Finally Breaks the Memory Wall

Today's Presenters

John Overton
CEO, Kove

Narendra Narang
Global Chief Architect, Red Hat TME

Chalapathy Neti
Head of AI CoE, Swift

Achieve More™

KOVE™

Red Hat

KOVE™

If Memory Serves Me Correctly….

Creating a “Constellation” of Memory Nodes

An Elastically Scalable Architecture for the Future of Distributed Computing

John Overton, CEO, Kove

Nita Chilapathy, Head AI, SWIFT

Narendra Narang – Global Chief Architect, Red Hat TME

Today’s Applications are Memory Hungry

.

1

MEMORY HUNGRY APPLICATIONS – AI Training, Financial Analysis, Blockchain, Graphics Rendering, Web3 and High Performance Computing (Scientific) applications all require large, dedicated memory capacities. In most cases, this excess memory simply sits in reserve, is fully powered on….and thus, expensive.

2

ACCESSIBILITY MEETS DEMAND – Hardware, GPUs and data are becoming increasingly accessible and available to a broader audience. All this demand comes at a cost - and again, requires dedicated memory.

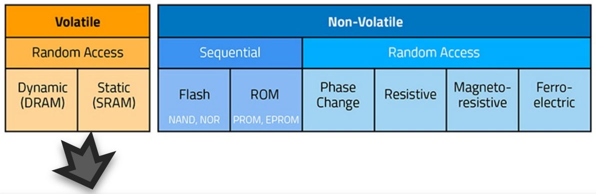
3

SECURITY & GOVERNANCE – Increasingly available access to large amounts of data makes the above applications easier to develop, but security and governance constraints force a hard look at where and how they are deployed. By being able to virtualize a shared pool of physical memory securely, these concerns can be addressed without purchasing more physical DRAM.

Memory and its Relation to Power

Volatile vs Non-Volatile

Memory Categories



Volatile		Non-Volatile					
Random Access		Sequential		Random Access			
Dynamic (DRAM)	Static (SRAM)	Flash NAND, NOR	ROM PROM, EPROM	Phase Change	Resistive	Magneto-resistive	Ferro-electric

Building for Peaks - Overall average utilization of DRAM 15-25%

Energy Consumption in a Server

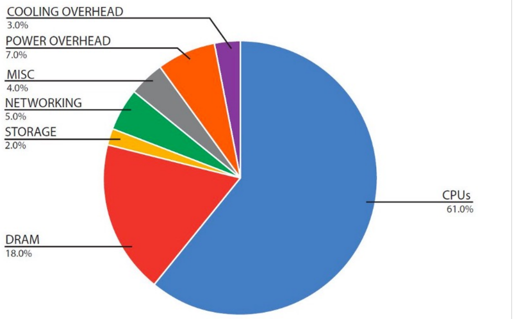
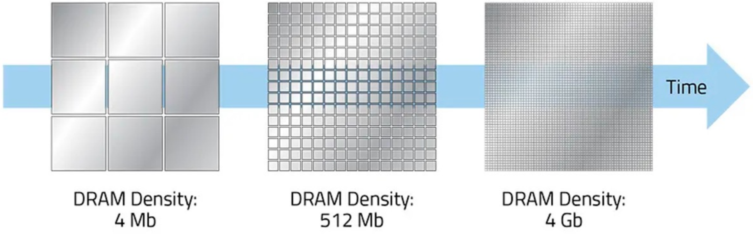


Figure 1.8: Approximate distribution of peak power usage by hardware subsystem in a modern data center using late 2017 generation servers. The figure assumes two-socket x86 servers and 12 DIMMs per server, and an average utilization of 80%.

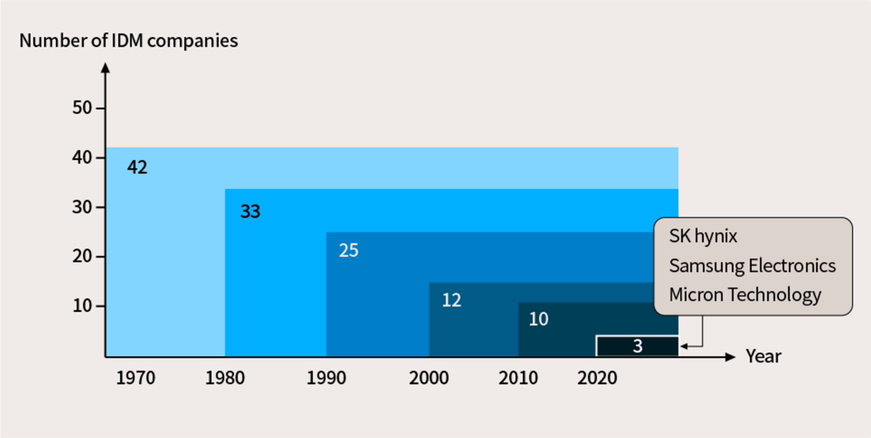
Why Focus on DRAM?

CPUs still consume maximum share, but we are **reaching the limits** of power optimization in CPUs.

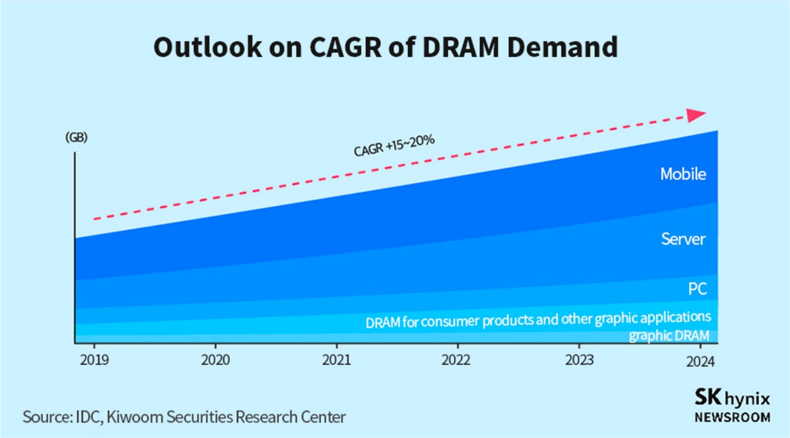
With increasing memory densities, DRAM becomes a hidden "**power hog**" in compute systems because of limitations.



Supply Chain Shrinking, Demand Growing, Costs Skyrocketing



https://news.skhynix.com/the-density-cost-and-marketing-of-semiconductor-memory/



Source: IDC, Kiwoom Securities Research Center

https://mediask.co.kr/90001-237

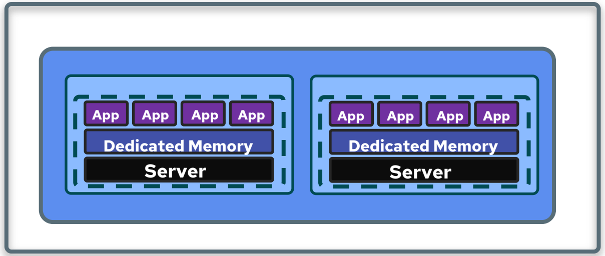
Memory demand for Mobile is exploding, which will require better management of Memory at the Edge of the Network (5G)

Re-thinking the Memory Model

What if we could simplify the model....and increase resiliency?

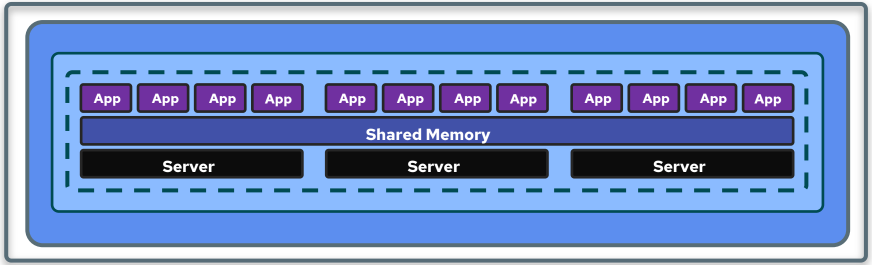
“Dedicated Memory Model”

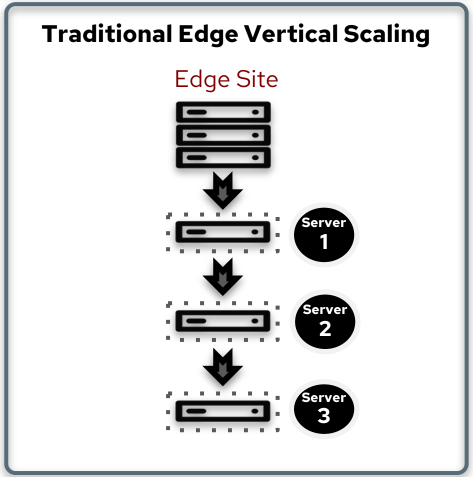
Stranded memory in each server



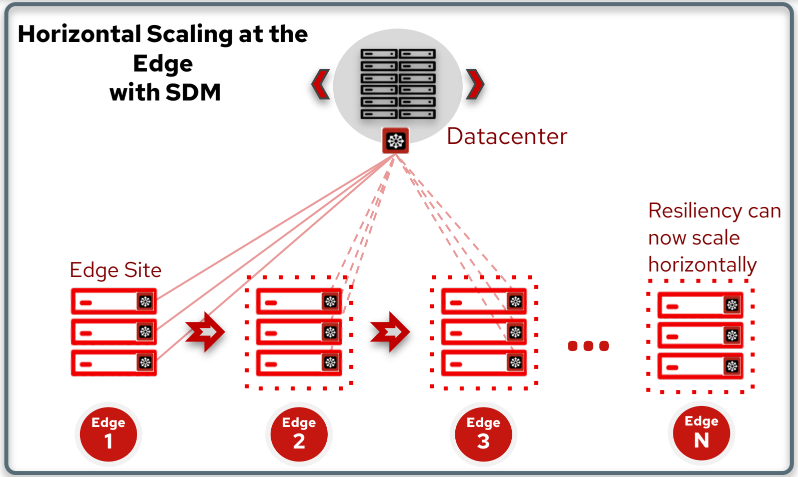
“Shared Memory Model”

No stranded memory





Traditional Edge Vertical Scaling

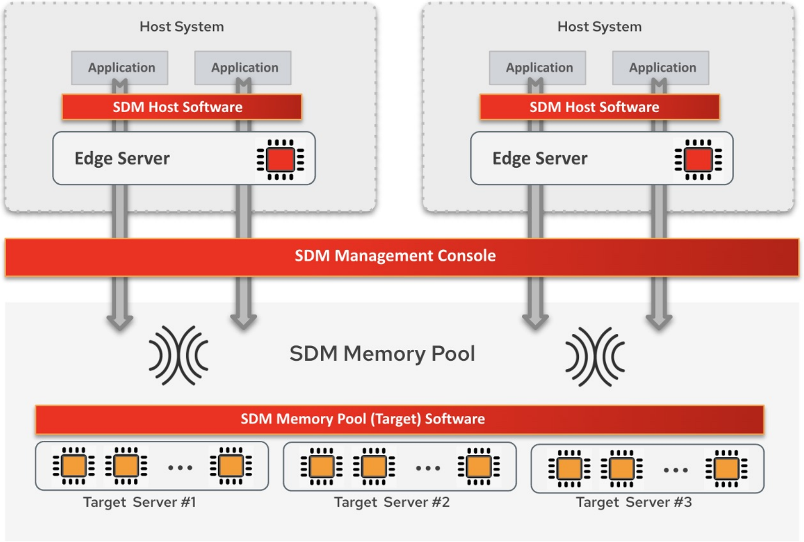


Horizontal Scaling at the Edge with SDM

Resiliency can now scale horizontally

Introducing Software-Defined-Memory (SDM)

Overcoming the Memory Wall!



Memory: One of the last frontiers

- Go beyond reducing CPU, I/O power consumption, and start efficiently using power for the DRAM.
- Clever algorithms can mask theoretical performance issues up to a limit.

How does it work?

- Software-Defined Memory (aka Virtualized Memory) enables individual servers to draw from a shared memory pool, receiving precisely the amount of memory needed, including amounts far larger than can be contained within a physical server.
- Memory pooling that results in more efficient utilization of resources.

Business Value

- Dynamically adjust for the working data set and burst on-demand to strategically utilize memory. Optimized CPU, memory, I/O, Network, and power consumption cost envelope.

Why SDM?

1 **SAVE TIME** – Without enough DRAM for your workload, you have to either partition it, or guess how much you need, and if you are wrong (as in an AI/ML training run) you lose all the time spent in running it and have to start over. Data scientist time is quite expensive…. Kove:SDM™ gives you enough virtual DRAM for your application, with no code modification.

2 **SAVE POWER** – No need to add additional memory per server in reserve, that needs to be powered on while waiting for utilization. All existing system memory is shared across every instance, driving efficiency and power savings.

3 **SAVE COST** – Enables economies of scale, as many data centers and Edge environments need to overprovision their memory systems. Kove:SDM™ eliminates unused memory in the Core and Edge, as all memory instances are fully utilized.

4 **REDUCE COMPLEXITY** – Partitioning workloads is complicated, and inferring results from smaller data sets is also complicated and inaccurate. With SDM, you can run your workload on the full, actual data set, in memory, with no code modification.

5 **ENHANCED SECURITY** – Kove SDM enables an empty “zero” memory instance for each application. Standard memory has other applications and data queued inside the same physical memory instance, which can be accessed by malicious actors. Kove disables any possibility of this happening with a “zero” instance – no other data is stored in the assigned memory instance, outside of what is required for that single application’s requirements.

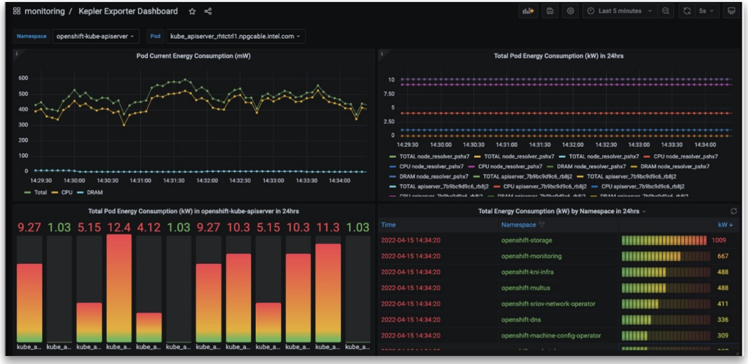
Red Hat

Power Reduction Potential 10%–40%

Combining Intel’s Kepler and Kove’s Software Defined Memory (SDM)

Intel CNCF Kepler Initiatives (20-30%)

(P-State Frequency/C-State Voltage)

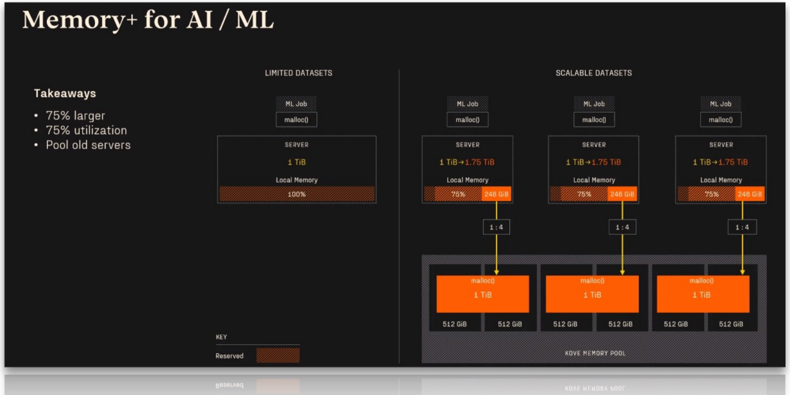

monitoring / Kepler Exporter Dashboard
Last 5 minutes
5s
Namespace
openshift-kube-apiserver
Pod
kube_apiserver_rhtctrl1.npgcable.intel.com
Pod Current Energy Consumption (mW)
600
500
400
300
200
100
0
14:29:30
14:30:00
14:30:30
14:31:00
14:31:30
14:32:00
14:32:30
14:33:00
14:33:30
14:34:00
Total
CPU
DRAM
Total Pod Energy Consumption (kW) in 24hrs
10
7.50
5
2.50
0
TOTAL node_resolver_pshx7
TOTAL node_resolver_pshx7
TOTAL node_resolver_pshx7
CPU node_resolver_pshx7
CPU node_resolver_pshx7
CPU node_resolver_pshx7
DRAM node_resolver_pshx7
DRAM node_resolver_pshx7
DRAM node_resolver_pshx7
TOTAL apiserver_7b9bc9d9c6_rb8j2
TOTAL apiserver_7b9bc9d9c6_rb8j2
TOTAL apiserver_7b9bc9d9c6_rb8j2
CPU apiserver_7b9bc9d9c6_rb8j2
CPU apiserver_7b9bc9d9c6_rb8j2
Total Pod Energy Consumption (kW) in openshift-kube-apiserver in 24hrs
9.27
1.03
5.15
12.4
4.12
1.03
9.27
10.3
5.15
10.3
11.3
1.03
kube_a...

Total Energy Consumption (kW) by Namespace in 24hrs

Time	Namespace	kW
2022-04-15 14:34:20	openshift-storage	1009
2022-04-15 14:34:20	openshift-monitoring	667
2022-04-15 14:34:20	openshift-kni-infra	488
2022-04-15 14:34:20	openshift-multus	488
2022-04-15 14:34:20	openshift-sriov-network-operator	411
2022-04-15 14:34:20	openshift-dns	336
2022-04-15 14:34:20	openshift-machine-config-operator	309

Software Defined Memory (5-15%)

(Share Memory Pool; Memory Frequency)



Memory+ for AI / ML

Takeaways

- 75% larger
- 75% utilization
- Pool old servers

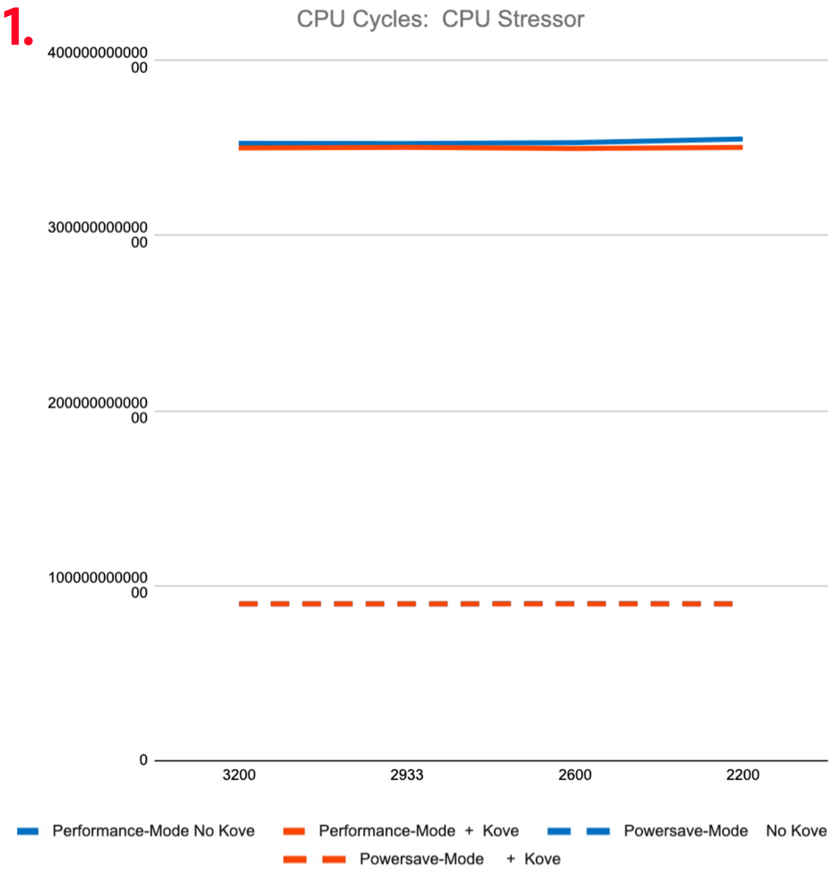
Key Drivers (SDM):

- Burst on-demand to dynamically adjust for the working data set
- Strategically utilize memory for CPU, memory and I/O bound workloads
- Optimized CPU, memory, I/O, Network, and power consumption cost envelope.

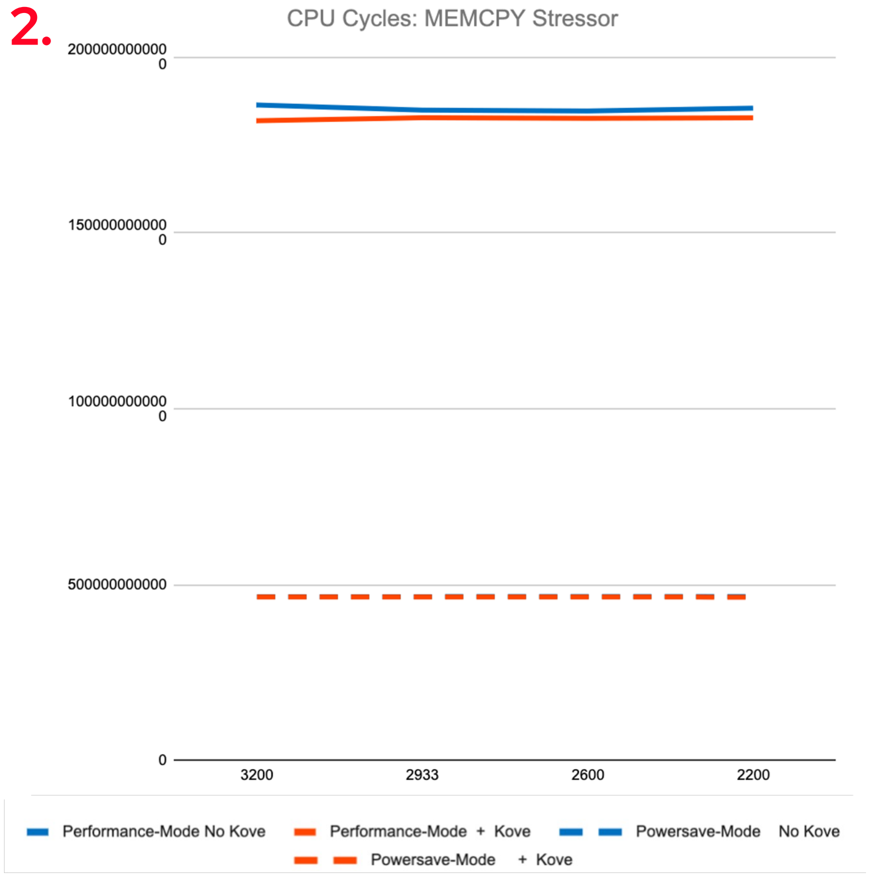
Minimal Impact on Memory Performance

CONFIDENTIAL – under NDA

1.



2.

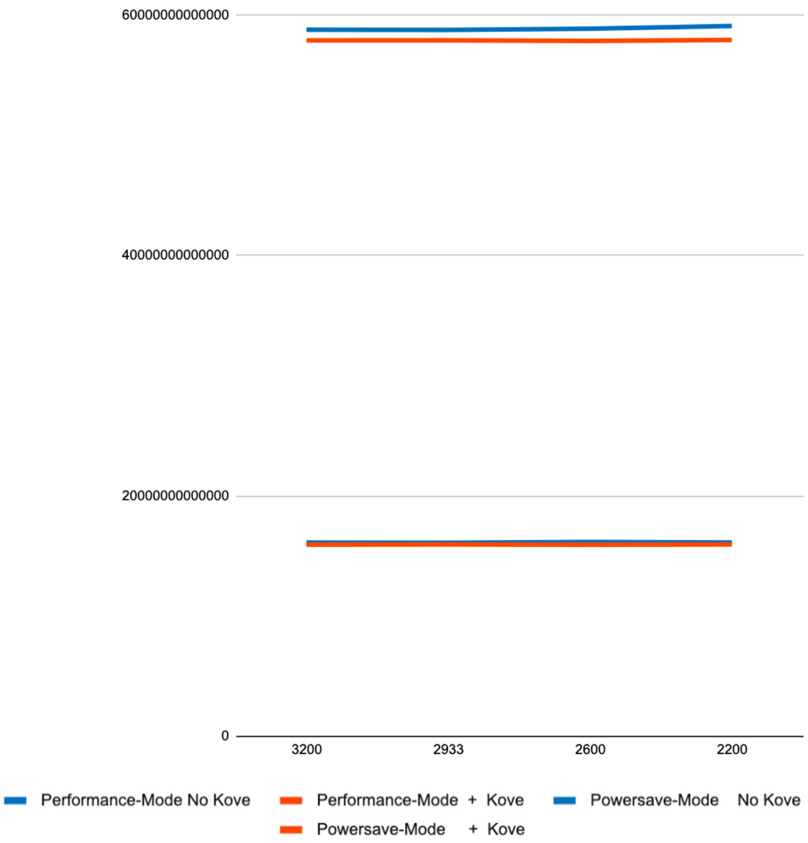


Memory Performance Can Even Be Faster

CONFIDENTIAL – under NDA

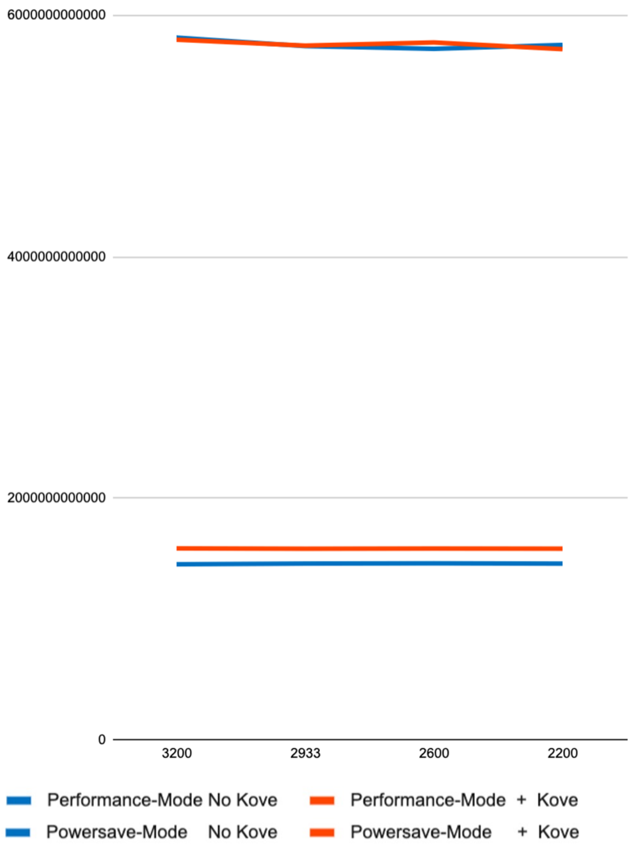
3.

Instructions: CPU Stressor



4.

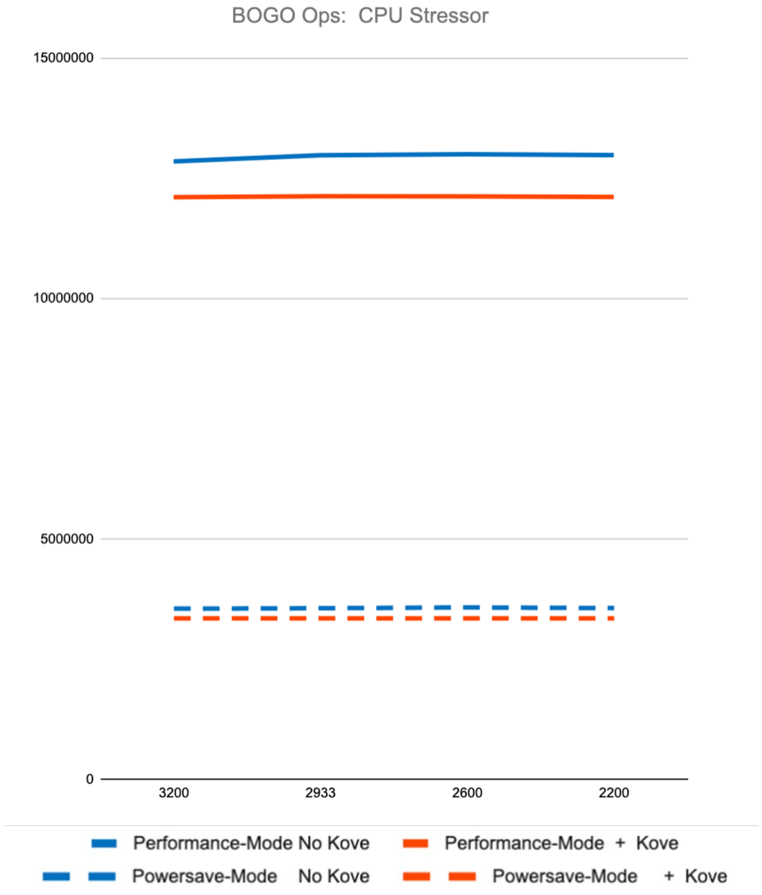
Instructions: MEMCPY Stressor



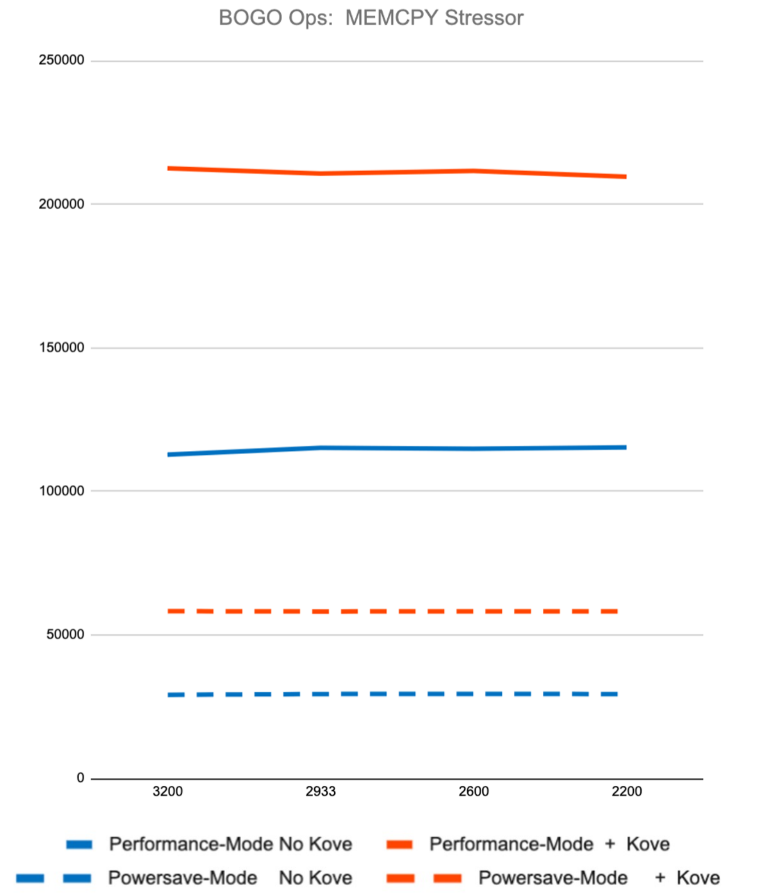
Significant Improvement in Memory Efficiency

CONFIDENTIAL – under NDA

5.



6.



Harnessing software-defined memory for near real-time fraud detection

Dr. Chalapathy Neti, Head of AI CoE, Swift

Swift

Presentation to MEMCON
March 24

Confidentiality: **Restricted**

March 24
Harnessing software-defined memory for near real-time fraud detection

Fraud in domestic and cross-border payments is growing annually

Financial institutions are investing significantly to safeguard the financial system

$485 billion

Annual cost of fraud

Nasdaq Verafin | 2024 Global Financial Crime Report

$202 billion

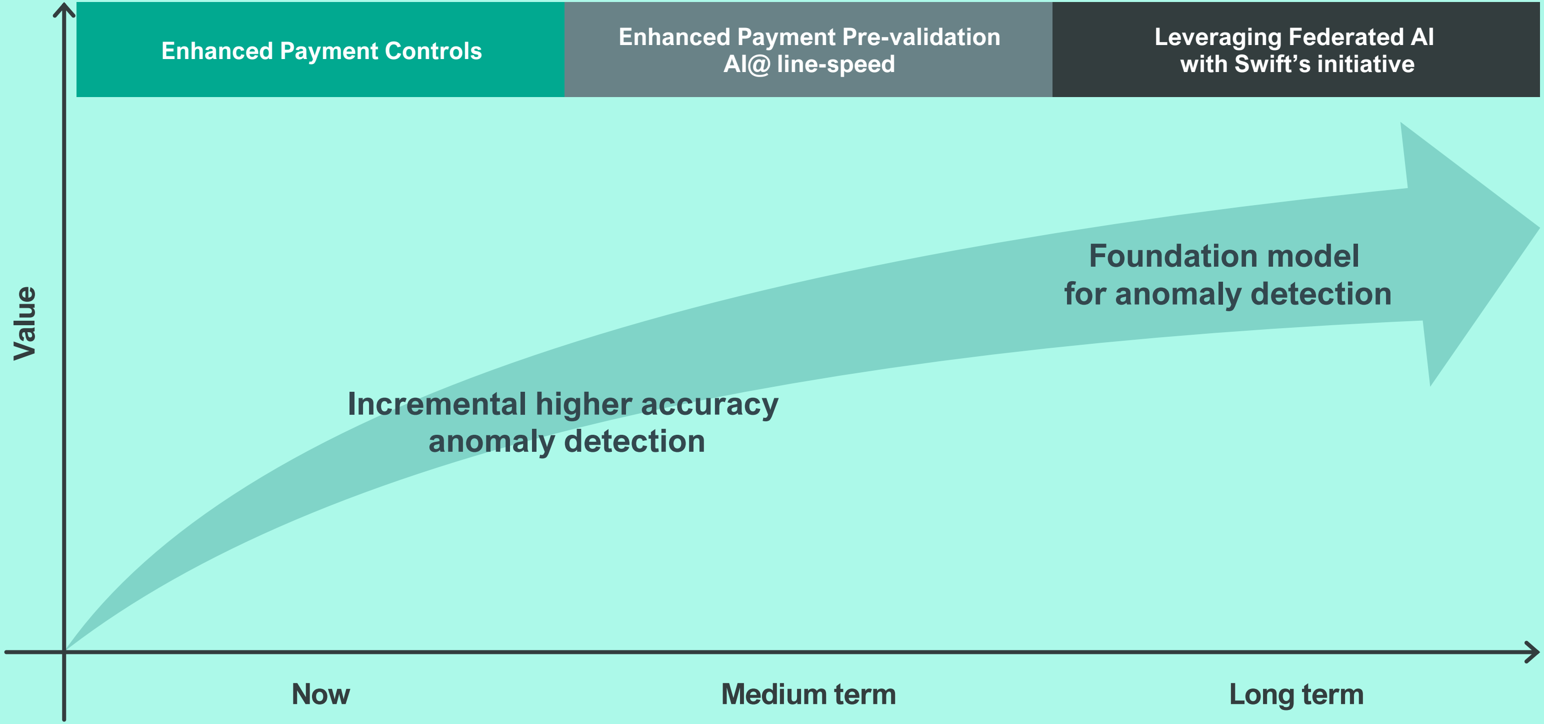
Annual cost of compliance

True Cost Of Financial Crime Compliance Study, 2023[1]

[1] A Forrester Consulting paper commissioned by LexisNexis® Risk Solutions, September 2023

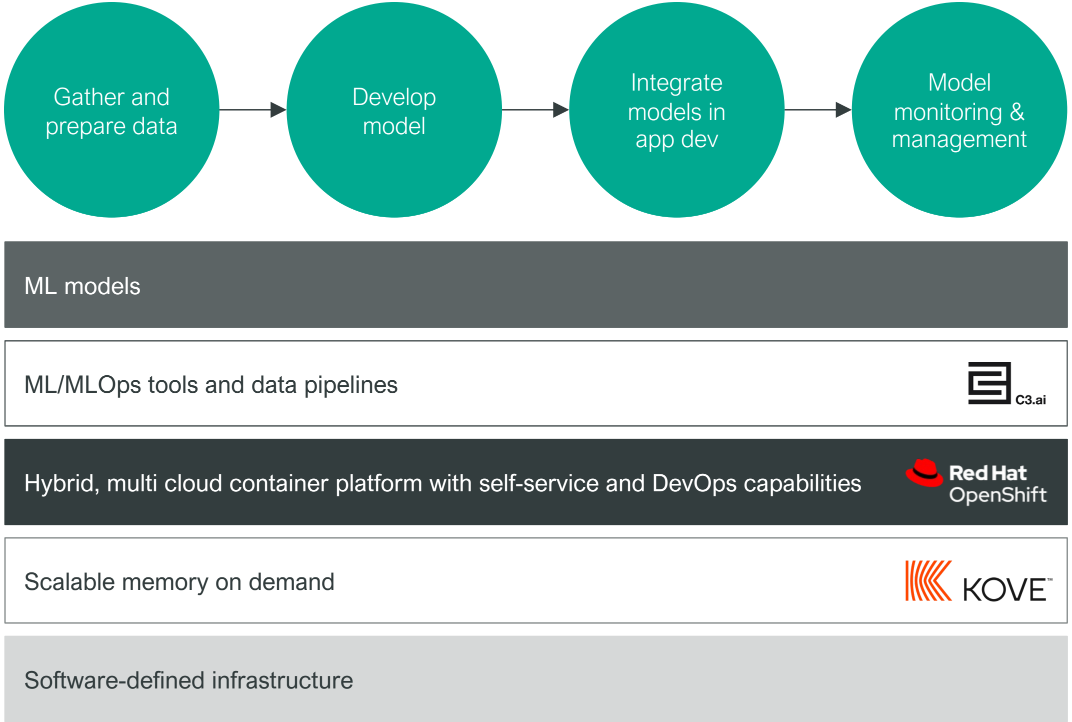
March 24
Harnessing software-defined memory for near real-time fraud detection

Swift, the largest cross-border financial messaging provider, is building a foundation model for anomaly detection in payments to enable near real-time, highly accurate transaction monitoring


Enhanced Payment Controls
Enhanced Payment Pre-validation AI@ line-speed
Leveraging Federated AI with Swift's initiative
Value
Foundation model for anomaly detection
Incremental higher accuracy anomaly detection
Now
Medium term
Long term

Partnering with Kove, Red Hat and C3.ai, Swift has developed an enterprise scalable AI platform to bring artificial intelligence applications to the +11,500 endpoints connected to its worldwide network

March 24
Harnessing software-defined memory for near real-time fraud detection



Gather and prepare data → Develop model → Integrate models in app dev → Model monitoring & management

ML models

ML/MLOps tools and data pipelines — C3.ai

Hybrid, multi cloud container platform with self-service and DevOps capabilities — Red Hat OpenShift

Scalable memory on demand — KOVE™

Software-defined infrastructure

"Leveraging Kove's software-defined memory capabilities and Red Hat's OpenShift containers on bare metal, we achieve an unprecedented **60x performance scalability** improvement compared to virtual machines on a hypervisor architecture, all at a competitive **commodity hardware price point**"

Swift

Swift is a global member-owned cooperative and the world's leading provider of secure financial messaging services.

We provide our community with a platform for messaging, standards for communicating and we offer products and services to facilitate access and integration; identification, analysis and regulatory compliance.

Our messaging platform, products and services connect more than 11,000 banking and securities organisations, market infrastructures and corporate customers in more than 200 countries and territories. Whilst Swift does not hold funds or manage accounts on behalf of customers, we enable our global community of users to communicate securely, exchanging standardised financial messages in a reliable way, thereby facilitating global and local financial flows, and supporting trade and commerce all around the world.

www.swift.com

Dr. Chalapathy Neti
chalapathy.neti@swift.com
+1 347 416 4571

Our local office address
Street: 7 Times Square, Fl. 45
Town: New York City
Country: United States

Swift

Thank You

Narendra & Chalapathy

KOVE™

Q&A

Kove:SDM™ Available Now

Unlimited Flexibility

—

100X More Containers

—

Achieve More™

CLAIM YOUR EXCLUSIVE TRIAL OFFER

—

Scan the code above or visit **kove.com/memcon** to learn more about trying Kove:SDM™ for yourself, plus get access to white papers, product demos and other exclusive content.