Key Value SSD Explained – Concept, Device, System, and Standard

YANG SEOK KI

Samsung Electronics

SDC STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2017

Disclaimer

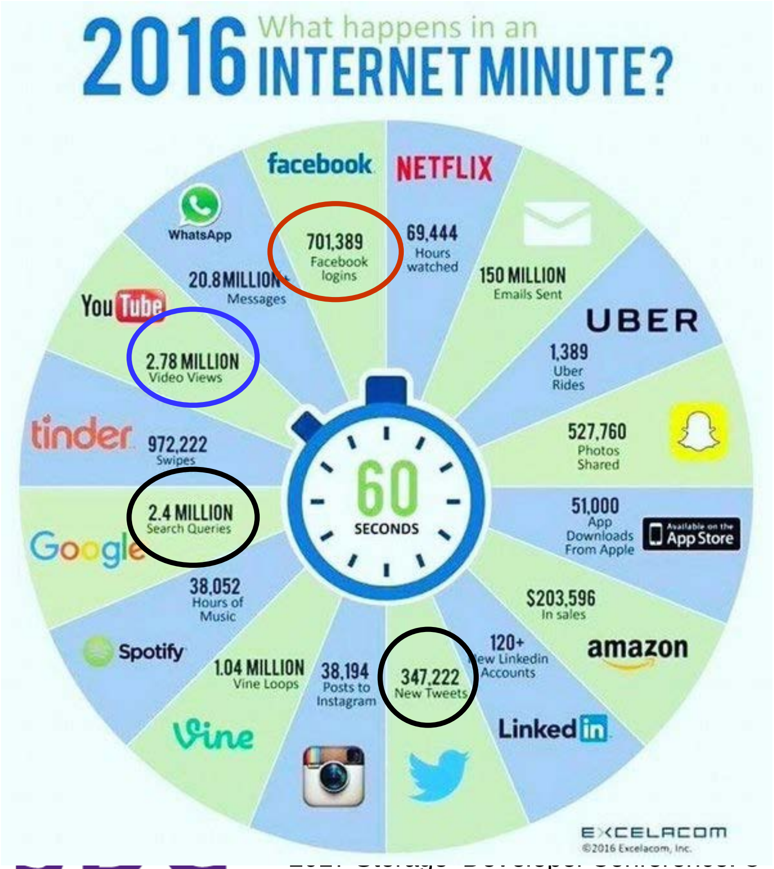
This presentation and/or accompanying oral statements by Samsung representatives collectively, the “Presentation”) is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, “Samsung”). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

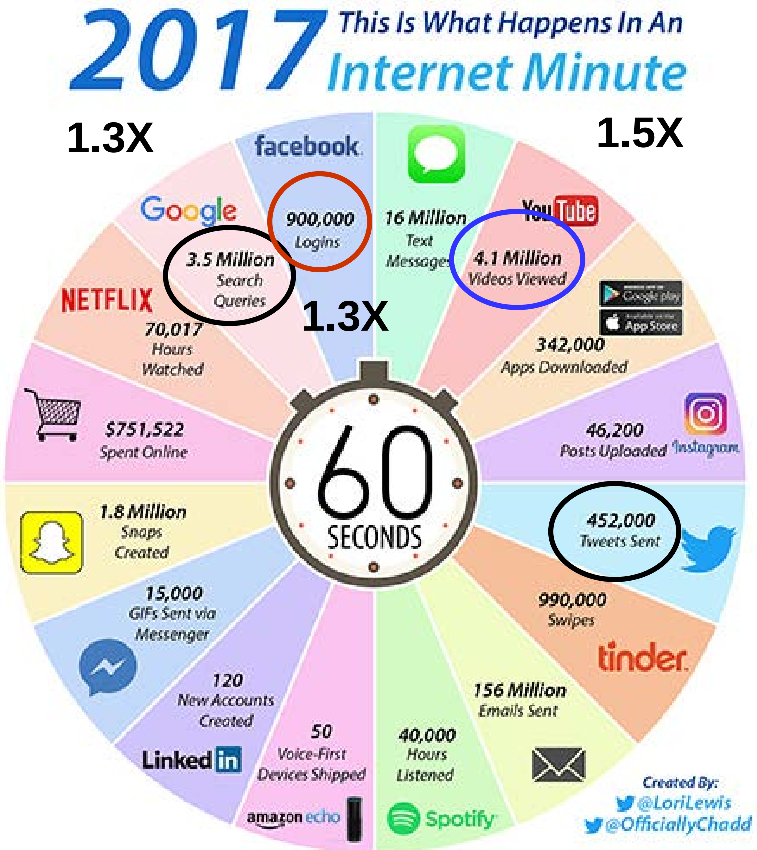
This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung’s intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.

Outline

- Background
- Concept
- Key Value SSD
- Ecosystem
- Standards
- Use Case Studies

What happens in an internet minute?





1.3X

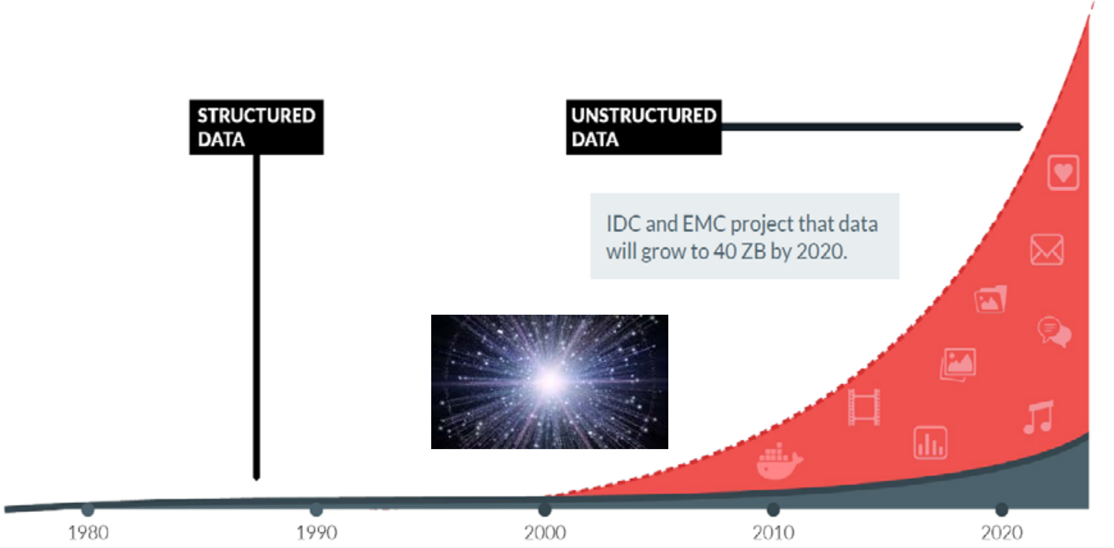
1.5X

1.3X

1.3X

BC/AD in IT

Source: Human Computer Interaction % Knowledge Discovery

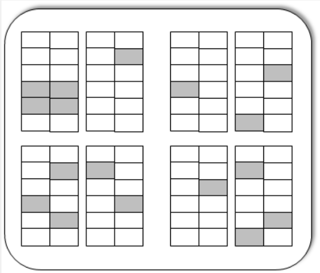


Before Cloud

Anno Datum

Everything is object!

Block



VS

Object

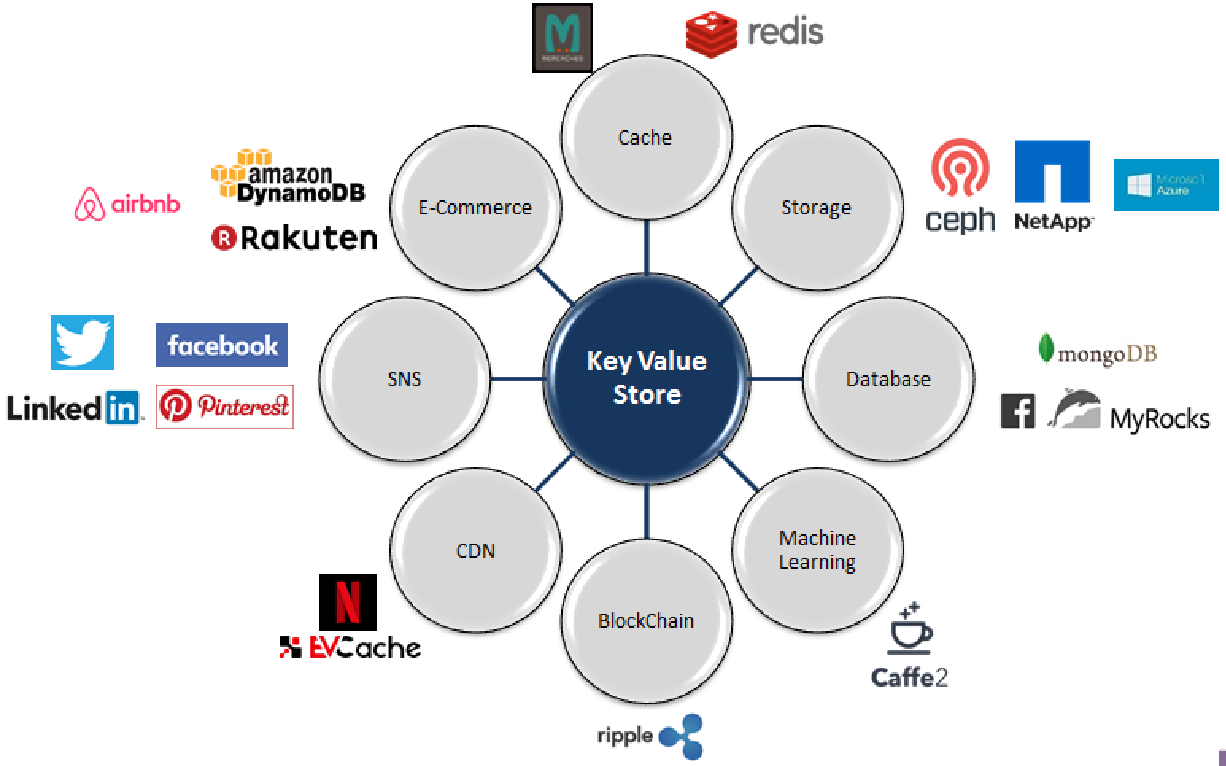
OSD Object Storage

ID | Attributes | User Data

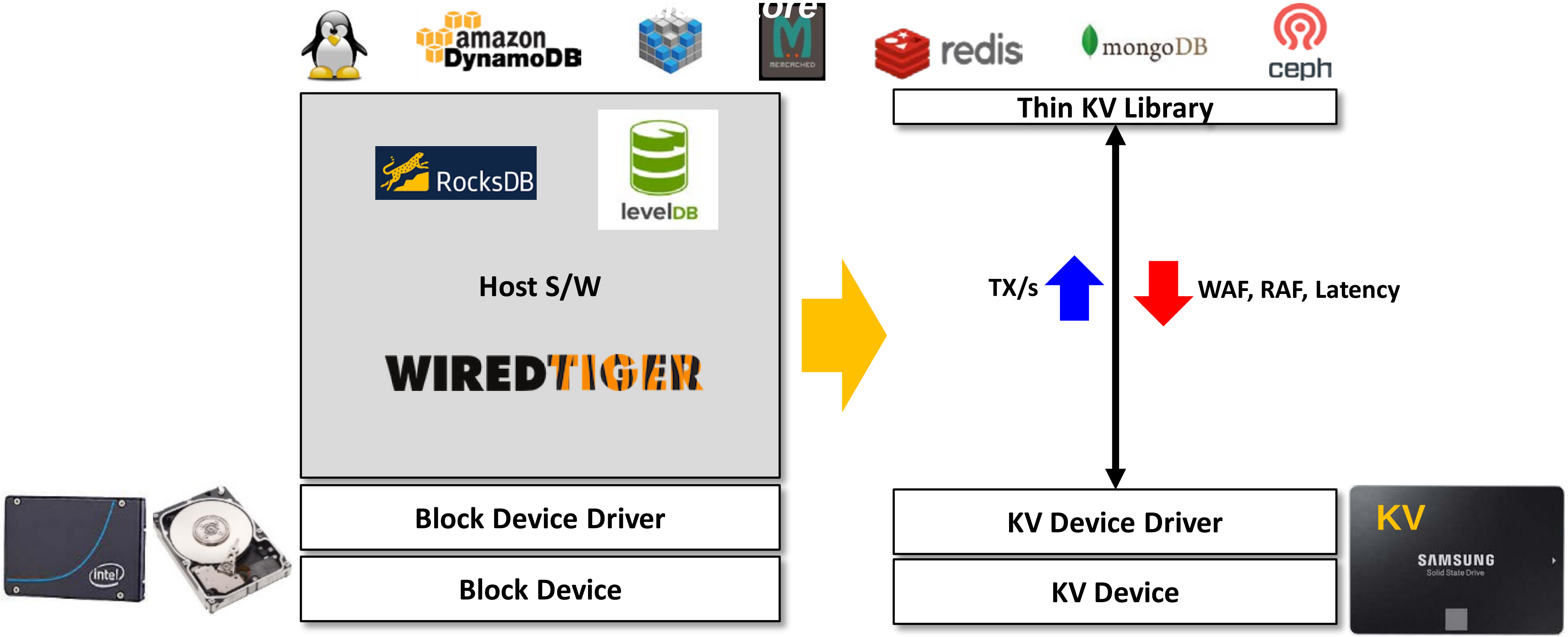
KV Storage

Key | Value

Key Value Stores are Common



Key Idea

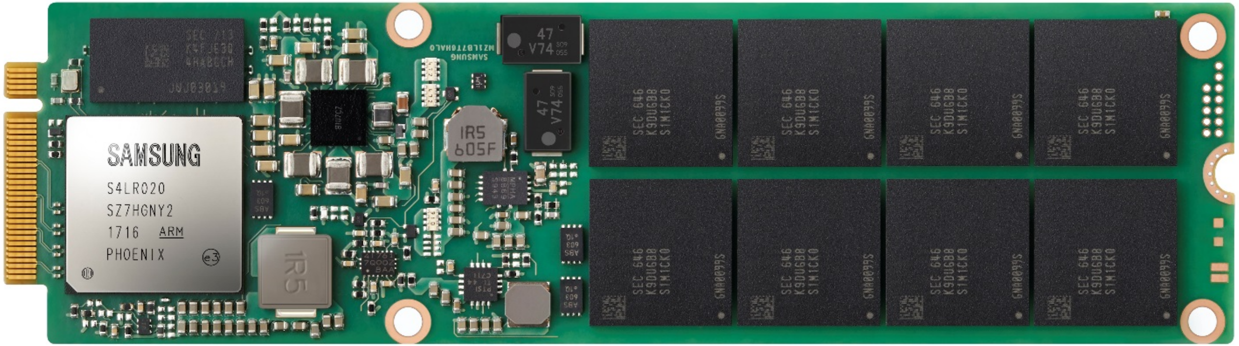


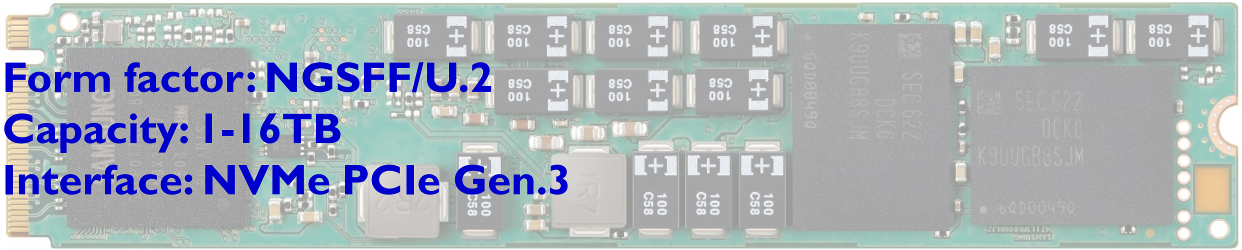
Traditional KV Store

KV Stacks

Samsung KV-PM983 Prototype

NGSFF KV SSD





Form factor: NGSFF/U.2
Capacity: 1-16TB
Interface: NVMe PCIe Gen.3

Key Value SSD is a Scalable Solution

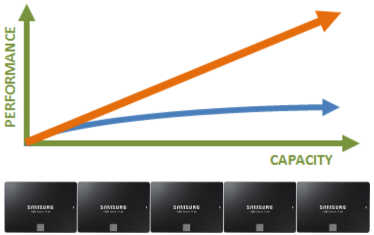
KV SSD
- Performance
- Capability

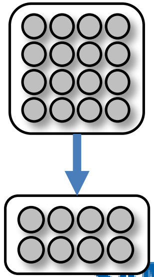
Scale-Up
- Capacity
- Performance

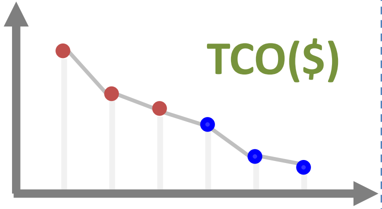
Scale-In
- CPU
- Server

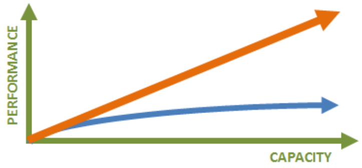
Scale-Down
- TCO
- Power

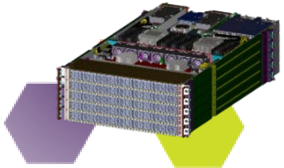
Scale-Out
- Capacity
- Performance



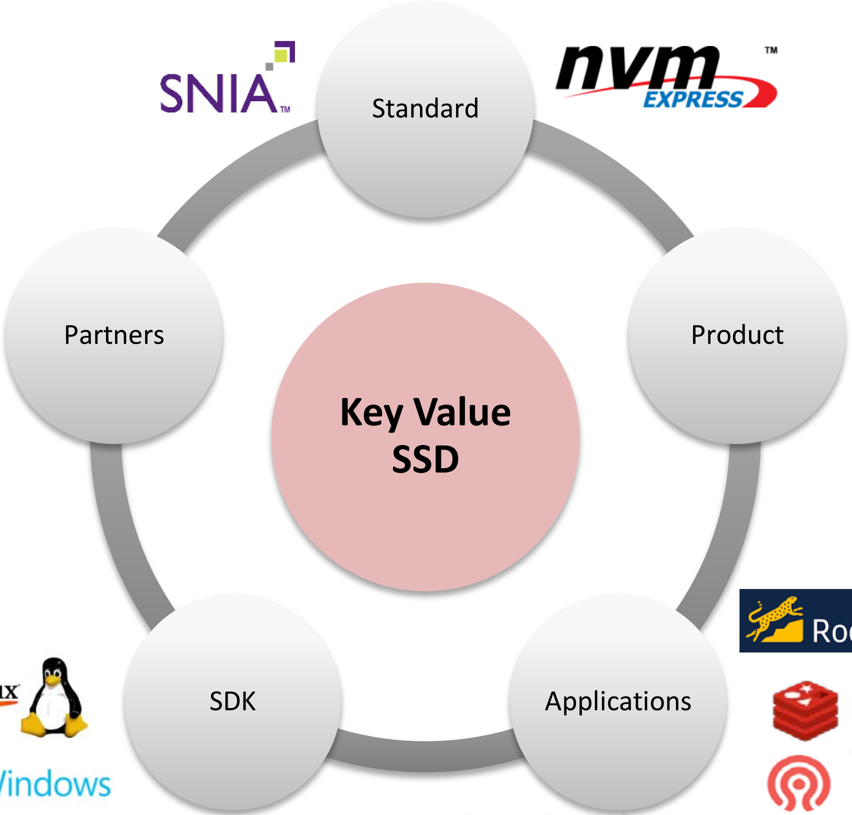


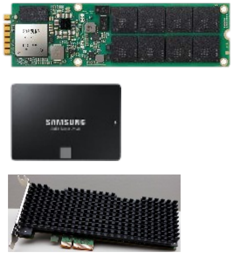






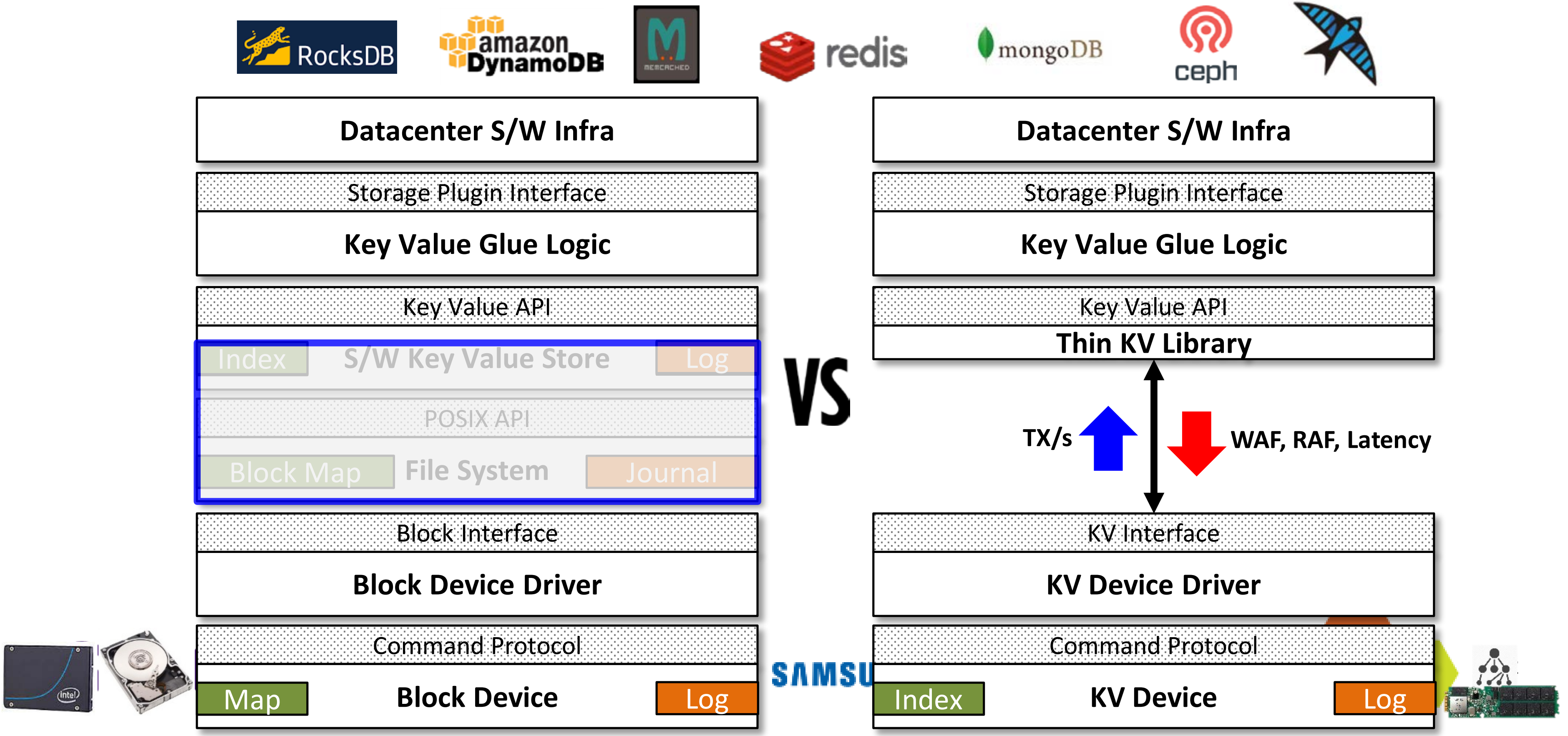
KV SSD Ecosystem





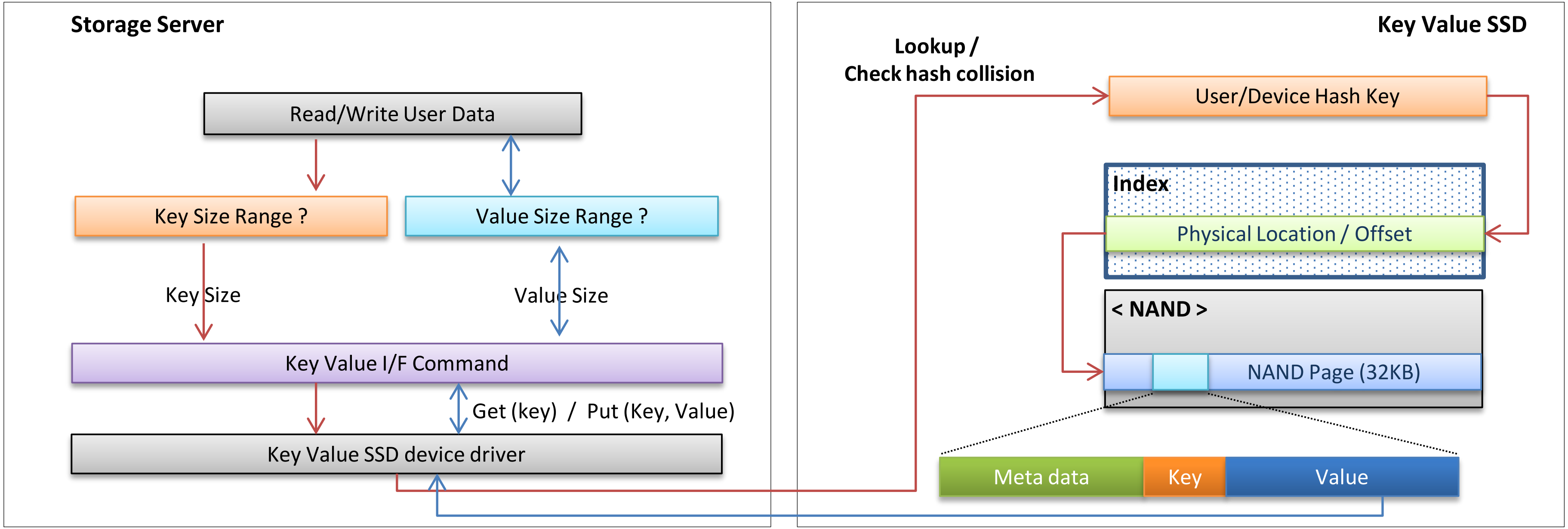
Key Value SW Stacks

- SSD with native key value interface through hardware software co-design



KV SSD Design Overview

- **SSD that supports native key value commands**



Key Value Software Development Stacks

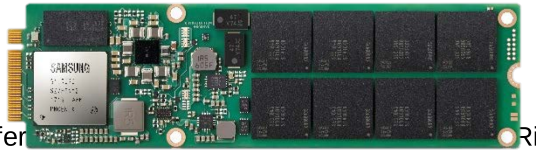
Key Value Library & Tools

- Cache
- AIO
- Multi-Queue
- Multi-Device
- Memory Manager
- Tools

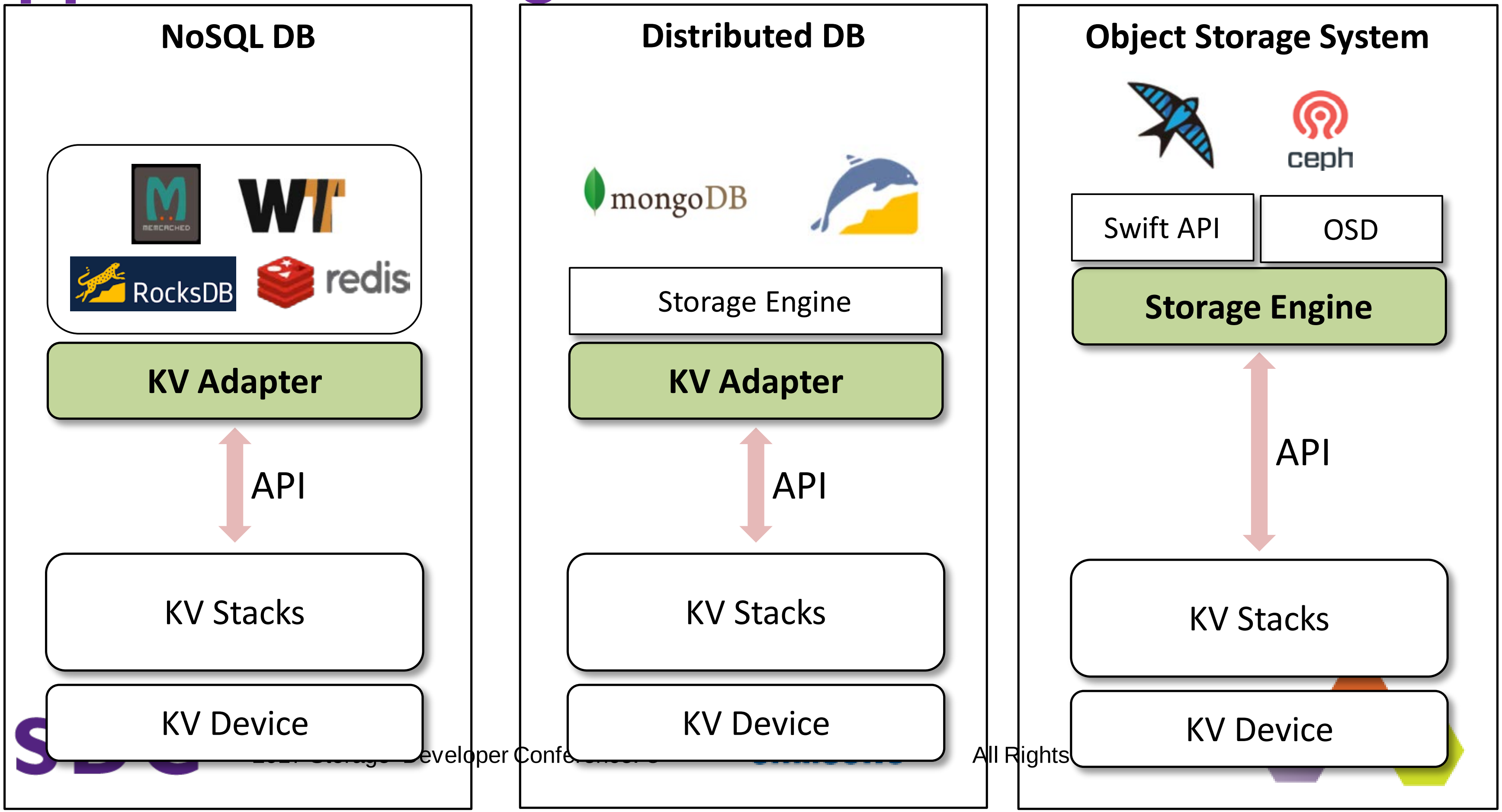
KV Abstract Device Interface (ADI)

- store/retrieve/delete/exist
- KV Pair
- namespace

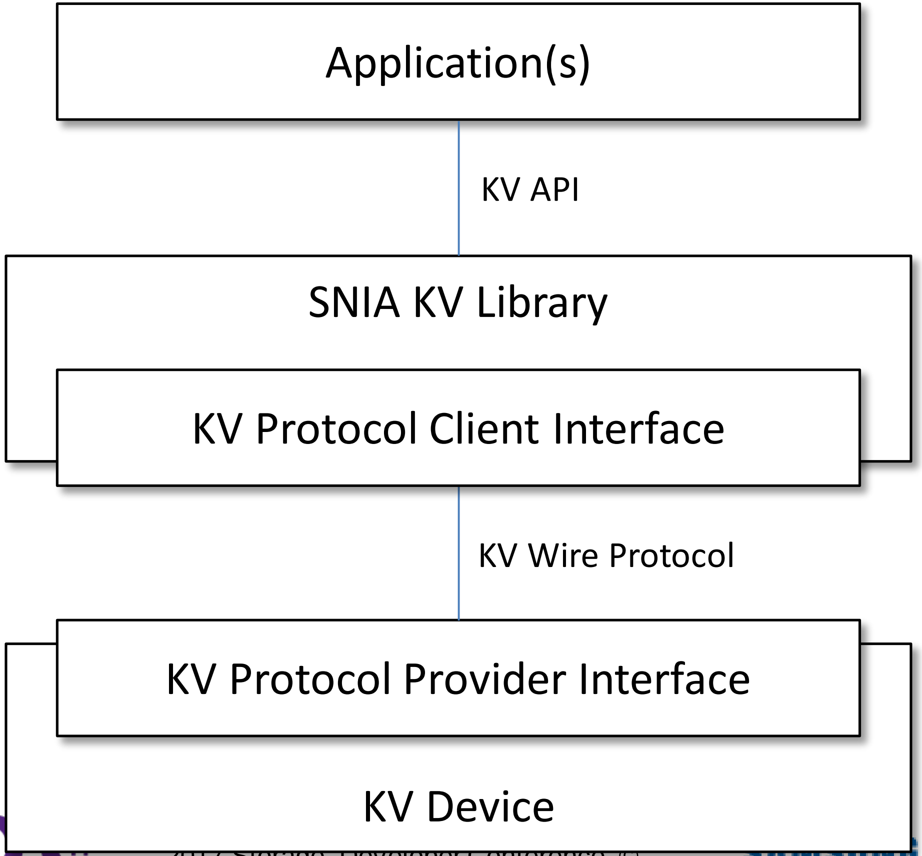
Linux Kernel Device Driver	Linux User-space Device Driver	Windows Device Driver



Application Integration with KV SSD



Key Value SSD Layers



e.g. SNIA KV API

e.g. C Library, Java, etc.
Written by vendors, open source, etc.

e.g. NVMe KV commands

e.g. KV SSD

Key Value SSD Standard Activities

- NVMe
  - Work on a technical proposal is being discussed by the NVMe working group
  - The group is defining the scope of the work
  - This will be a new device type
- SNIA
  - A proposal for a Key Value API has been submitted to the SNIA Object Drive Technical Working Group
  - Discussion on the minimum necessary commands to meet basic Key Value needs is progressing

Key Value, not Object Drive

- Both standards efforts are focused on Key Value SSD not Object Drive
  - Key Value is a means to submit a Key and put or get a Value
  - Object Drive would include more extensive commands to query the Key Value database

NVMe Extension for Key Value SSD

- **Defines a new device type for a Key Value device**
- **A controller performs either KV or traditional block storage commands**

New Key Value Commands	PUT	GET	DELETE	EXISTS
Existing Command Extension	Admin command	Identify commands for KV	Other non-block specific commands	

SNIA Key Value API

- The Key Value API (Application Programming Interface) has been presented to SNIA for consideration in the Object Drive Technical Working Group
- Defines a Tuple
  - Key
  - Value
- Defines KV specific constants
  - Max Key Length
  - Alignment Unit
- Key type supported
  - 4 byte fixed
  - 8 byte fixed
  - Variable length character string
  - Variable length binary string
- The API defines the calls that an application may make to the Key Value device interface
  - These calls are independent of any specific implementation
  - These calls support the basic commands proposed for the NVMe standard
    - Open/Close
    - Store/Retrieve
    - Exist
    - Delete
    - Containers/groups

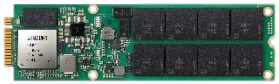
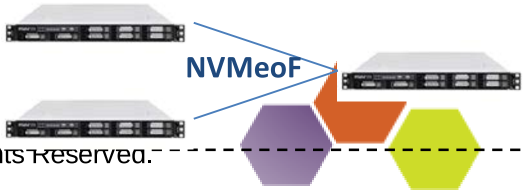
Call for Participation

- NVMe work is proceeding in the NVMe working group
  - www.nvmexpress.org
    - Contributors and Promoters have access to working proposals
- SNIA work is proceeding in SNIA Object Drive Technical Working group
  - www.snia.org
    - Members may join the Object Drive TWG and have access to working proposals

SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2017

Use Case Studies

Use Case Study

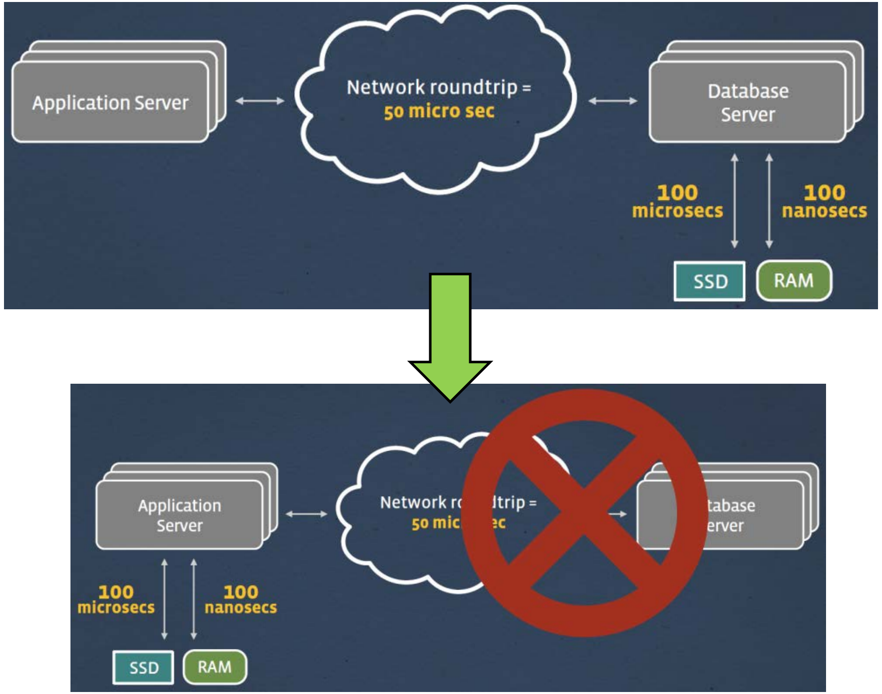
	Single	Scale-Up	Scale-Out
Benchmark	KVBench		
Key Value Store	RocksDB vs KV Stacks		
Device		 SAMSUNG	

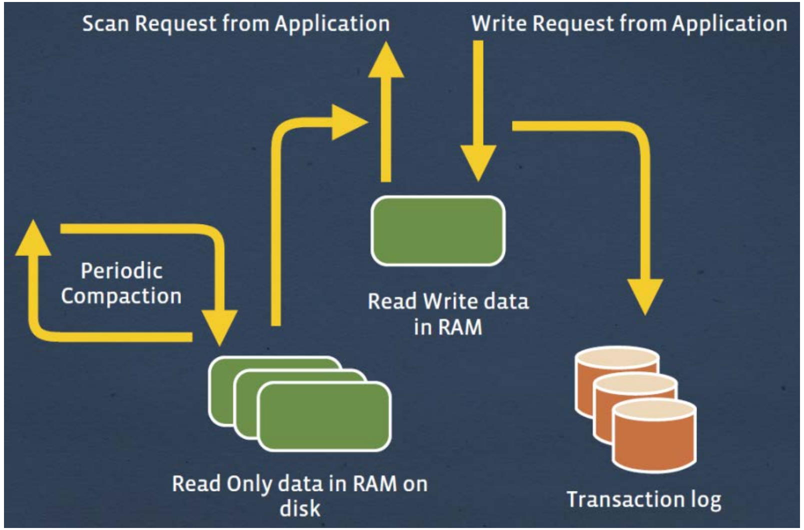
Single Component Performance: RocksDB vs. KV Stacks

- RocksDB
  - Originated by Facebook and Actively used in their infrastructure
  - Most popular embedded NoSQL database
  - Persistent Key-Value Store
  - Optimized for fast storage (e.g., SSD)
  - Uses Log Structured Merge Tree architecture

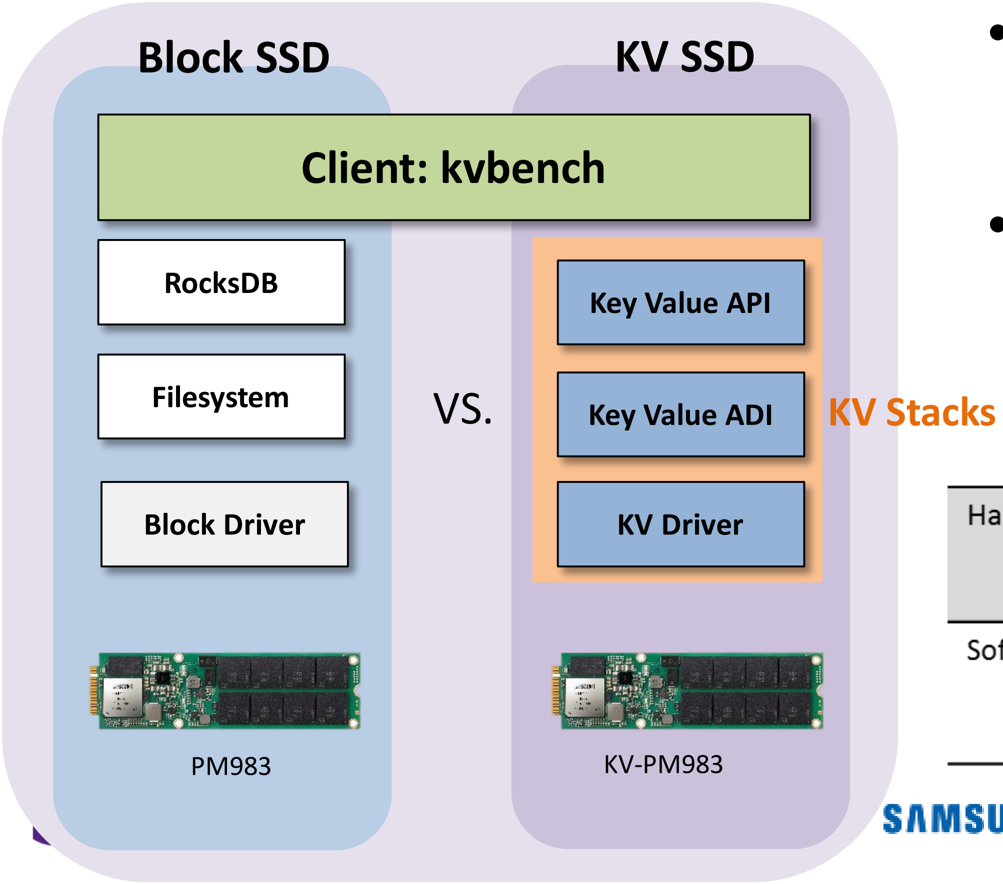
- KV Stacks on KV SSD
  - Benchmark tool directly operates on KV SSD through KV Stacks

RocksDB: Key Value Database





RocksDB vs. KV Stacks Performance Measurement

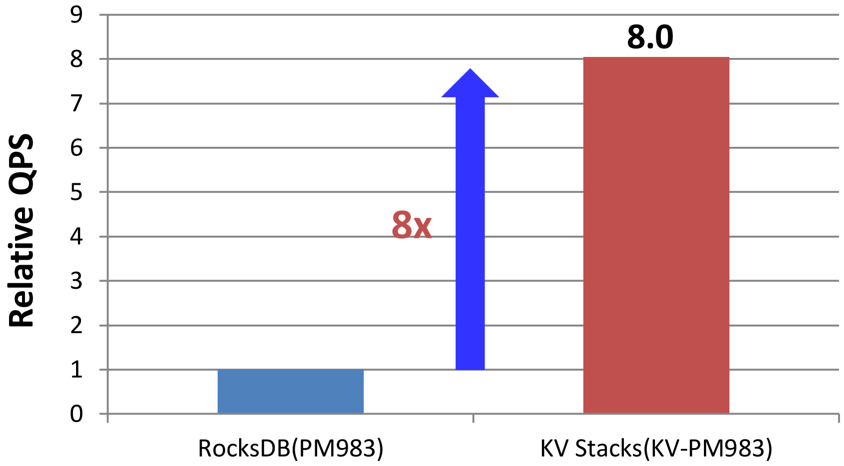


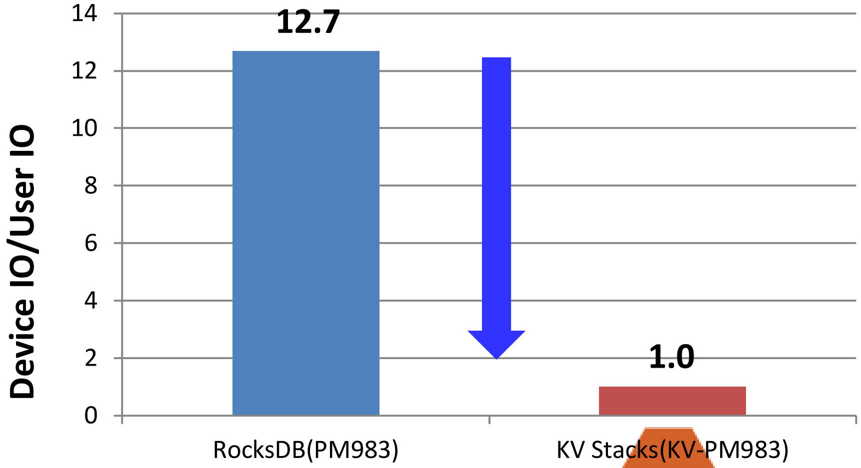
- **Better Performance**
  - Lean software stacks
  - Overhead moved to device
- **IO Efficiency**
  - Reduction of host traffic to devices

Hardware	Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz 96 GB RAM PM983(Block) & KV-PM983 SSD
Software	Ubuntu 16.04 RocksDB v5.0.2 on XFS 50M records, 16B Key, 4KB value

Performance: Random PUT

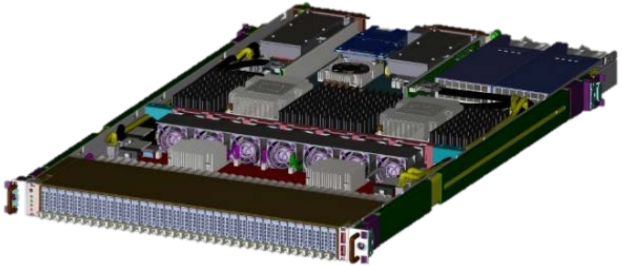
- **8x more QPS (Query Per Second) with KV Stacks than RocksDB on block SSD**
- **90+% less traffic goes from host to device with KV SSD than RocksDB on block device**



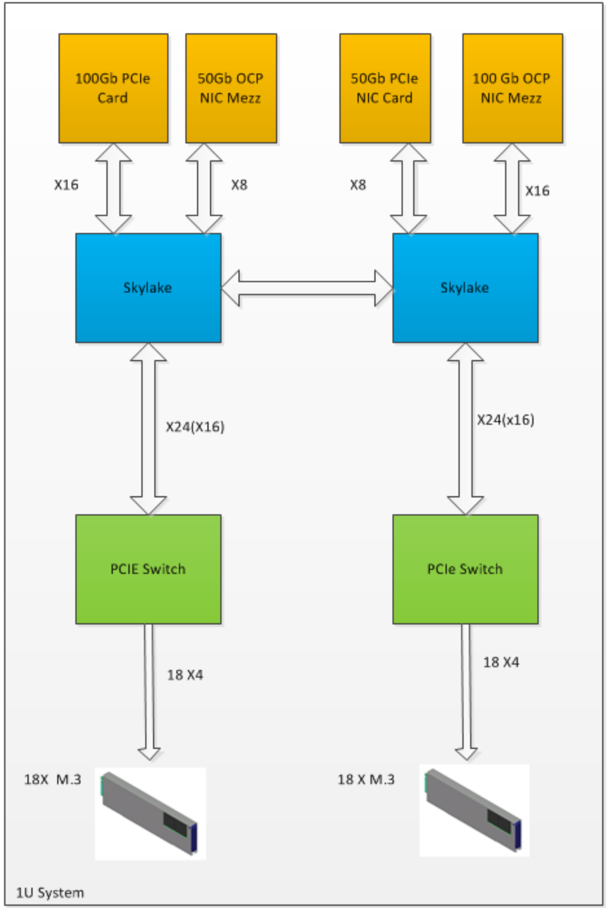


* Workload: 100% random put, 16 byte keys of random uniform distribution, 4KB-fixed values on single PM983 and KV-PM983 in a clean state

Testbed System for Scaling



CPU	Xeon 8160 CPU @2.10GHz 1-node 2-socket, 48-core 96-thread
# SSDs	36x 1TB

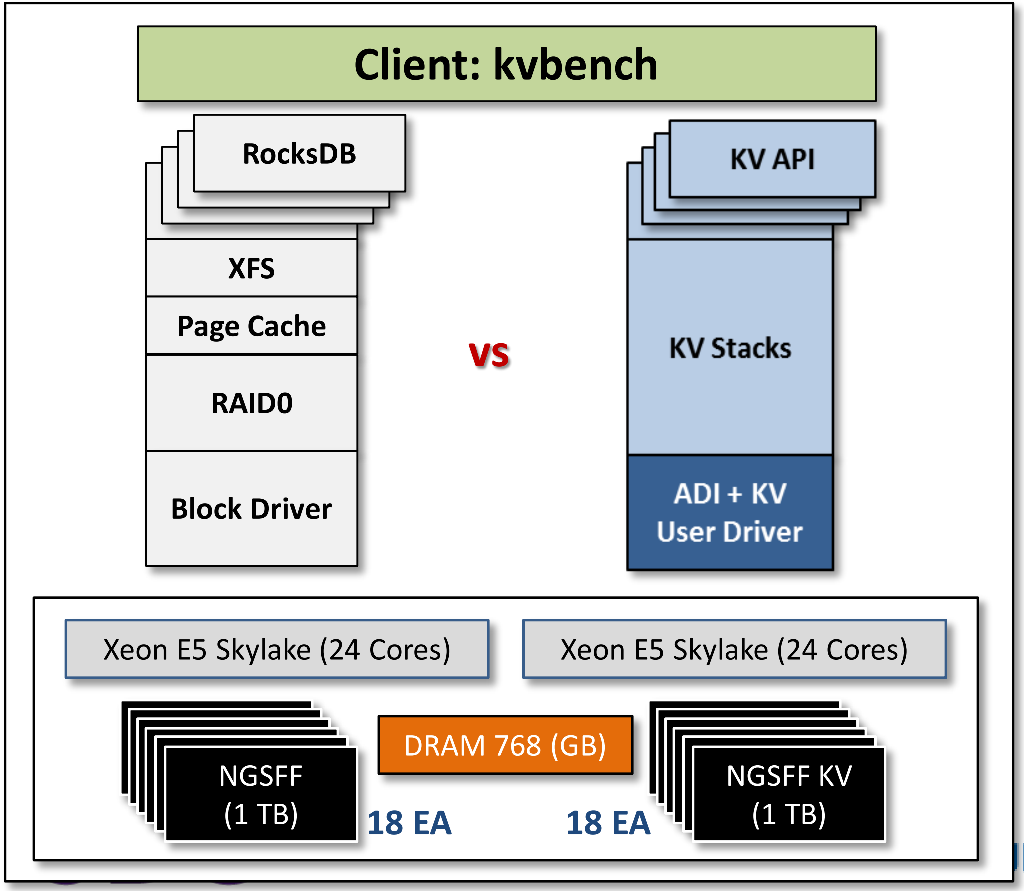


Full Box: 7.8 MIOPS

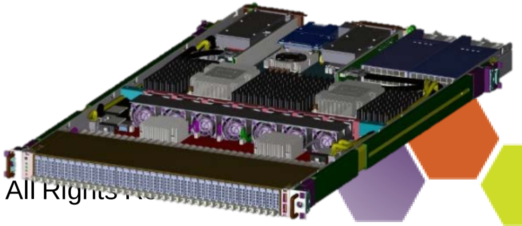
Fully Balanced Architecture
No PCIe bottle Neck

Fully utilize the Skylake CPU and PCIe switch bandwidth

Scale-Up Storage: RocksDB

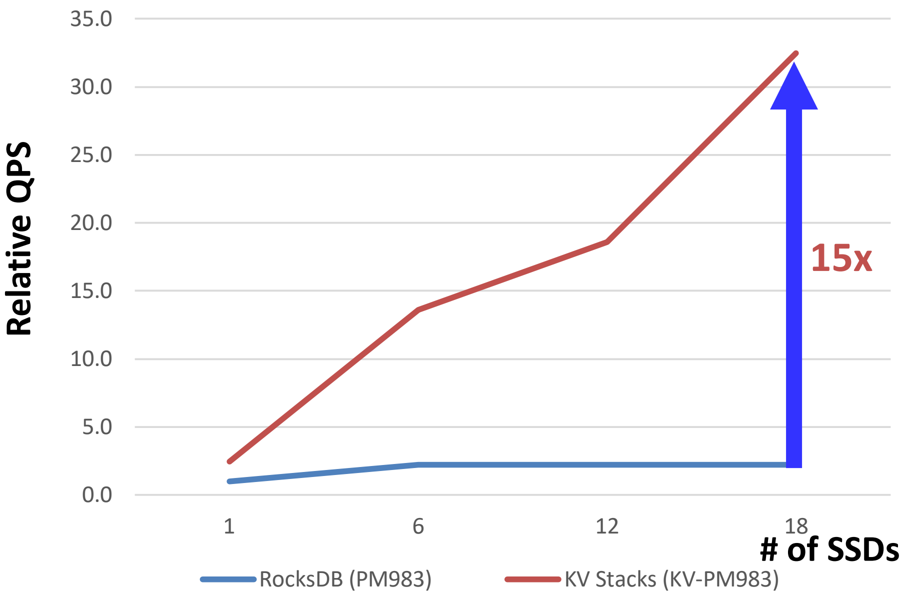


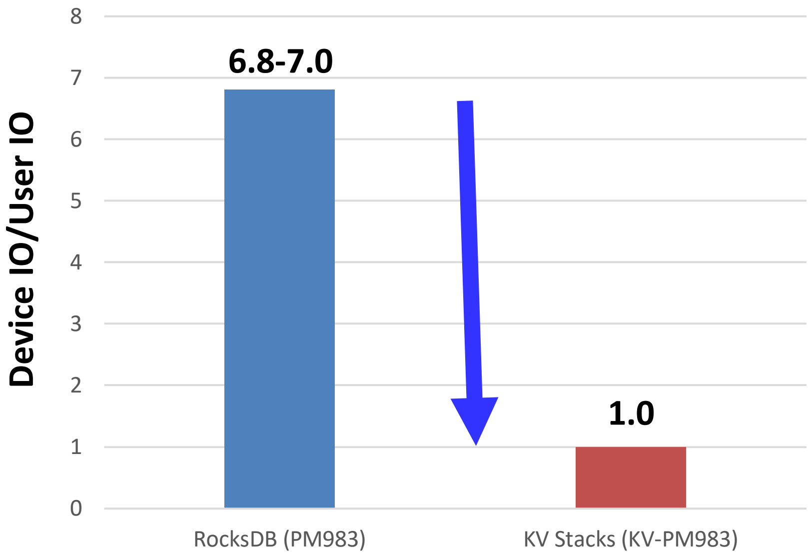
- **Linear Scaling**
  – More devices, more throughput and capacity
- **IO Efficiency**
  – Reduction of host traffics to devices
- **Less CPU utilization**
  – Small number of cores or less CPU utilization for performance



Scale-up Performance: Random Key PUT

- **15x IO performance over S/W key value store on block devices**

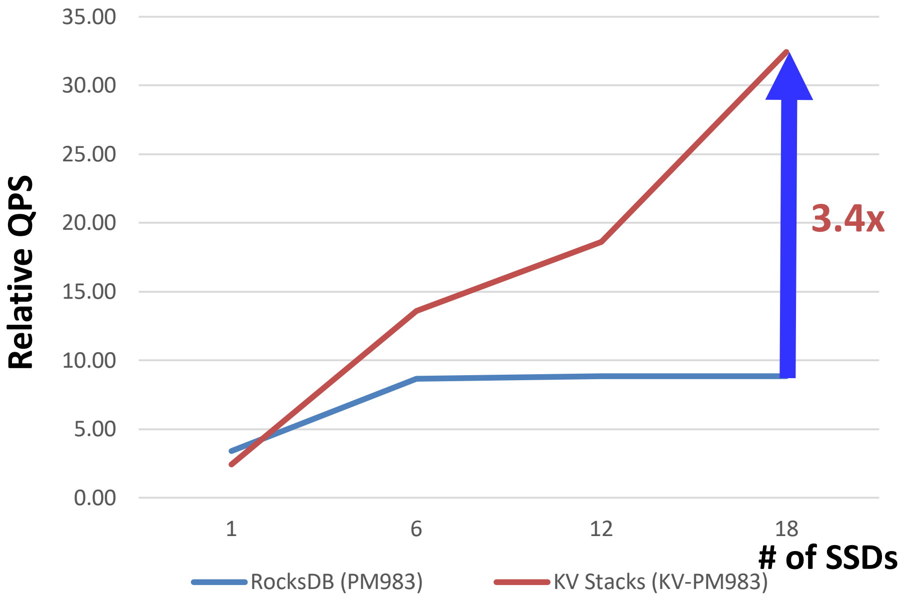


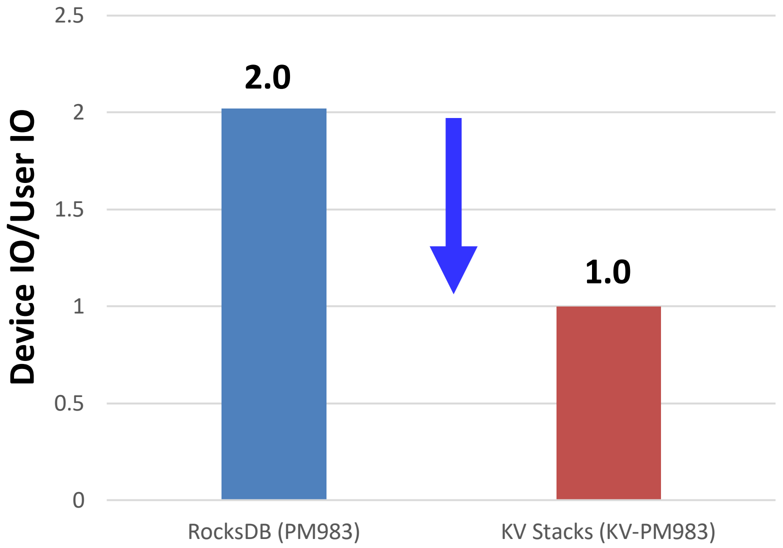


Relative performance to the maximum aggregate RocksDB random Put QPS for 1 SSD with a default configuration for 1 PM983 SSD in a clean state.
System: Ubuntu 16.04.2 LTS, , Ext4, RAID0 for block SSDs, Actual CPU utilization could be 70-90% at CPU saturation point.
Workload: 100% puts, 16 byte keys of random uniform distribution for RocksDB v. 5.0.2, 4KB-fixed values, 24 RocksDB Instances with 4 client threads, 50GB/Instance or 1.2TB Data is used

Scale-up Performance: Sequential Key PUT

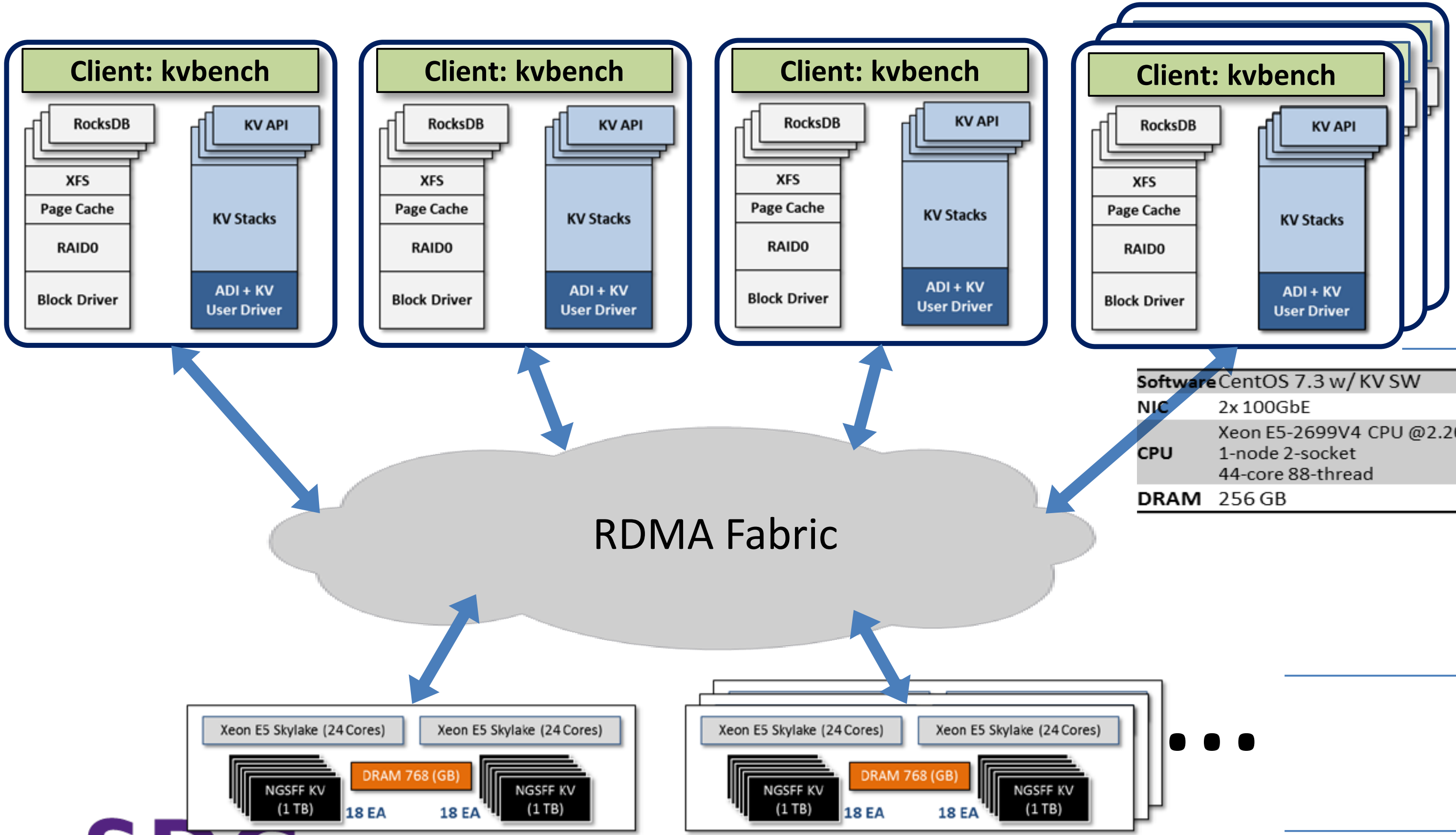
- **3.4x IO performance over S/W key value store on block devices**





Relative performance to the maximum aggregate RocksDB random Put QPS for 1 SSD with a default configuration for 1 PM983 SSD in a clean state.
System: Ubuntu 16.04.2 LTS, , Ext4, RAID0 for block SSDs, Actual CPU utilization could be 90% at CPU saturation point.
Workload: 100% puts, 16 byte keys of random uniform distribution for RocksDB v. 5.0.2, 4KB-fixed values, 36 RocksDB instances with 1 client thread, 34GB/Instance or 1.2TB Data is used

Scale-Out: RocksDB & KV Stacks Configuration



Software	CentOS 7.3 w/ KV SW
NIC	2x 100GbE
CPU	Xeon E5-2699V4 CPU @2.20GHz 1-node 2-socket 44-core 88-thread
DRAM	256 GB

CPU	Xeon 8160 CPU @2.10GHz 1-node 2-socket, 48-core 96-thread
# SSDs	36x 1TB

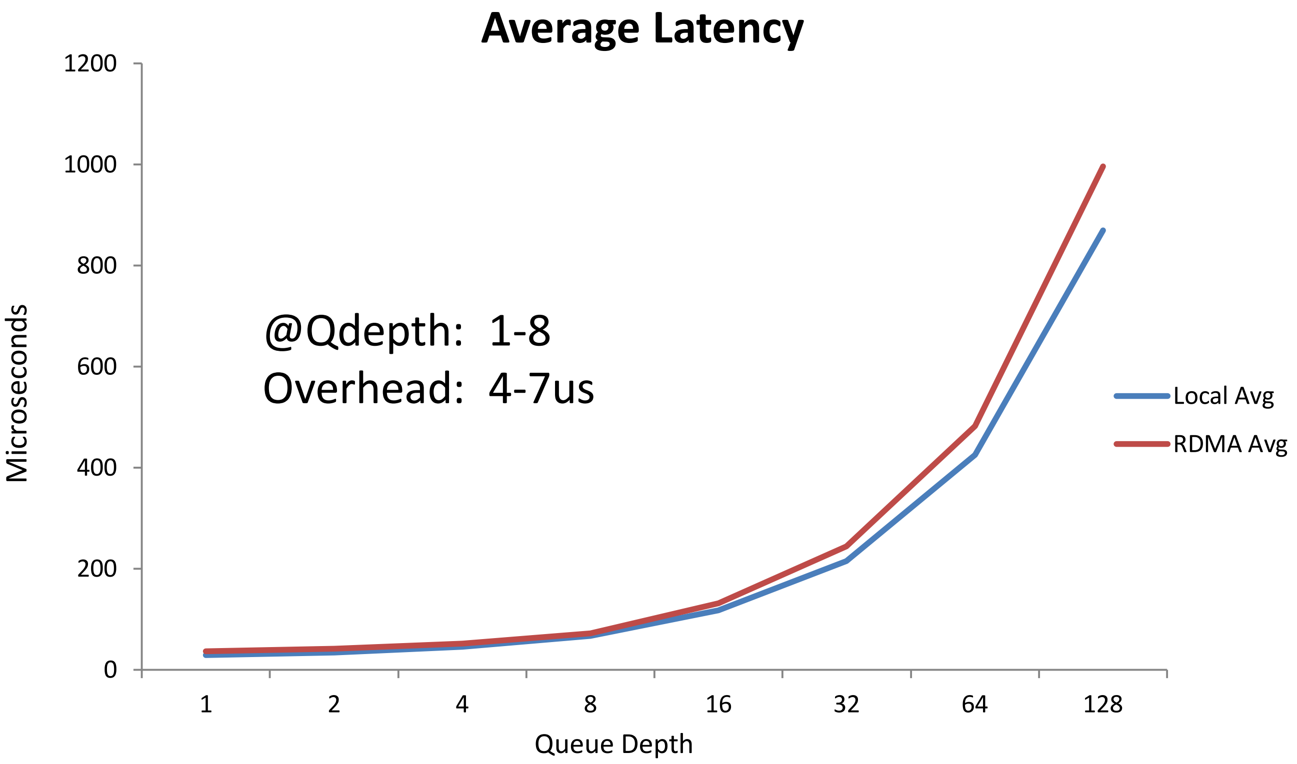
Software	CentOS 7.3 w/ KV Target
NICs	2x 100GbE + 2x 50GbE

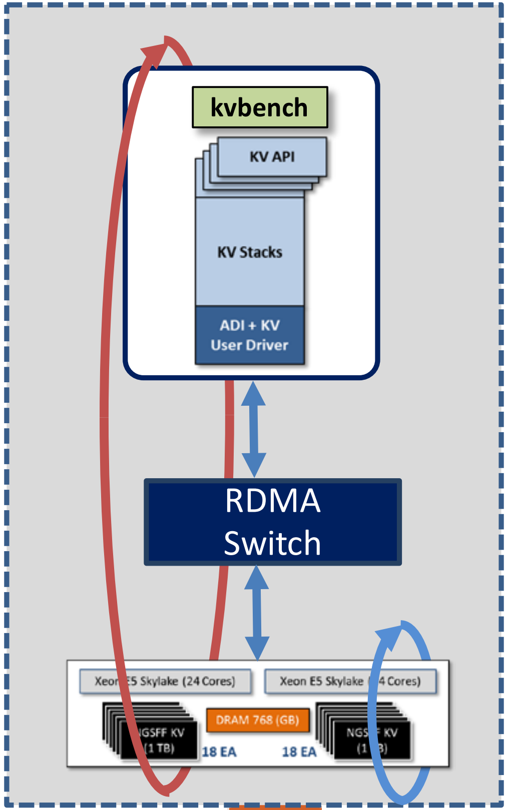
RocksDB vs KV Stacks

NVMeoF over RDMA

Mission Peak KV-PM983 SSDs

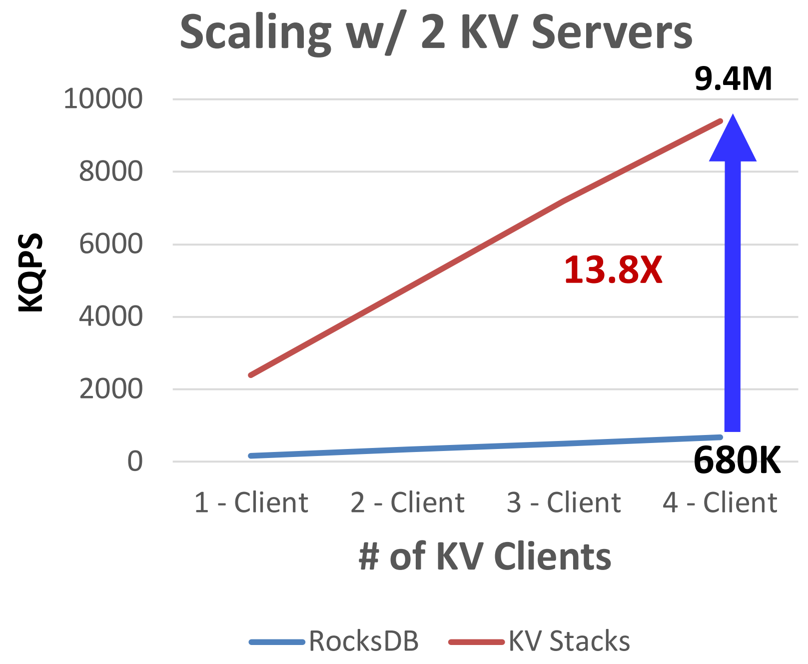
Local vs NVMeoF PUT Latency



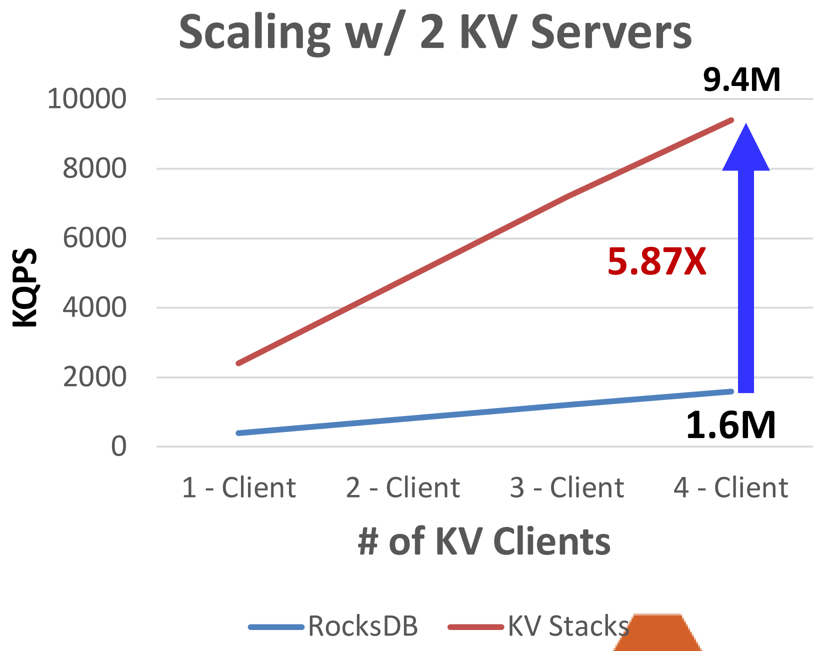


Performance and Capacity Scale-Out: PUT Throughput

Fill Random



Fill Sequential



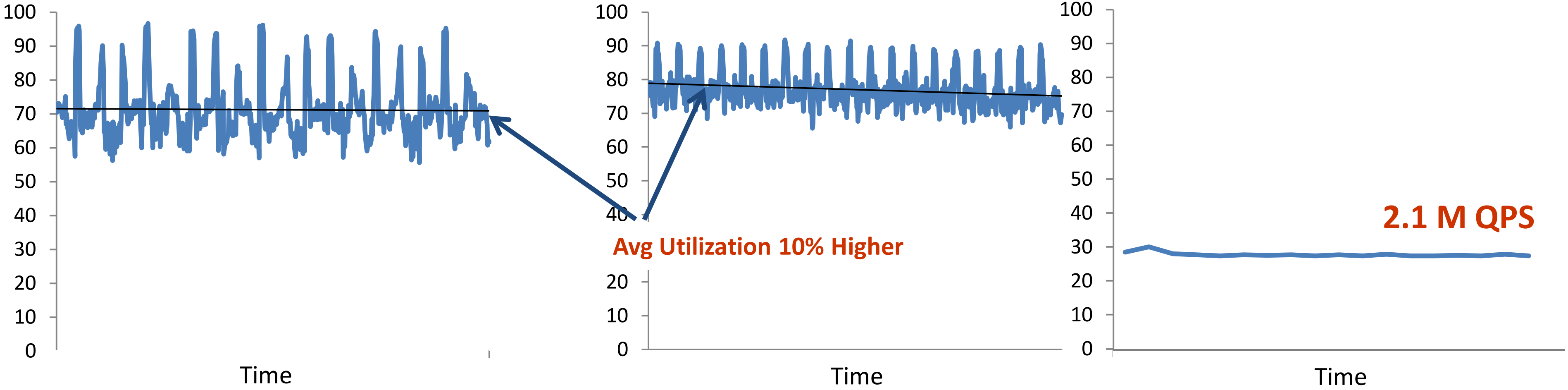
Client RocksDB: CentOS 7.3, Ext4, RAID0 for block SSDs,
Workload: 100% puts, 16 byte keys of random uniform distribution for RocksDB, 4KB-fixed values, 24 RocksDB instances with 8 client threads, 50GB/Instance or 1.2TB Data is used,
Client KV Stacks: CentOS 7.3, KV Load Generator, 100% 4K PUTs, 16 byte keys,
KV Server: Mission Peak w/ NVMeoF KV Target

CPU Utilization for Clients

Fill Random

Fill Sequential

KV Stacks

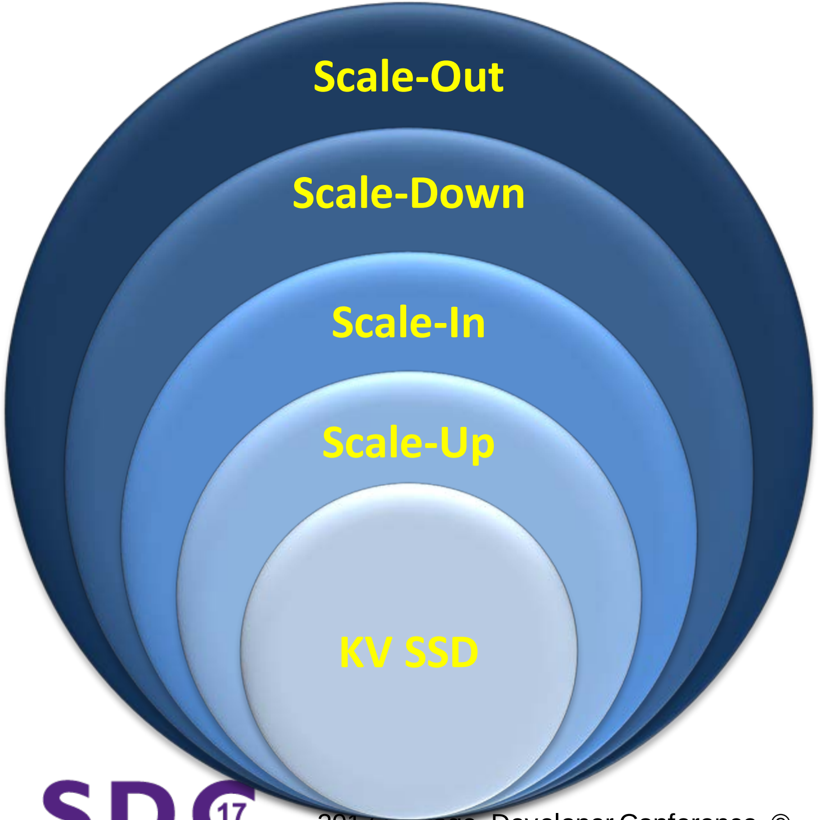


Avg 170K QPS@72% CPU

Avg 400K QPS@80% CPU

Avg 2.1M QPS@30% CPU

Conclusion



- Linear performance and capacity scaling
- TCO reduction
- CPU or server reduction
- Dense performance and capacity scaling
- Lean host software stacks

36

Questions?

kvssd@ssi.samsung.com