Decibel: Isolation and Sharing in Disaggregated Rack-Scale Storage

Mihir Nanavati, Jake Wires, and Andrew Warfield
(Coho Data and University of British Columbia)

WD Micron EMC² amazon web services™

Redundancy and high availability

Transparent device aggregation

Cluster-accessibility

Decoupling and migration

SEAGATE Microsoft Azure NetApp®

Splendid cheeses they were with a two hundred horse-power scent that might have been warranted to knock a man over at two hundred yards.

Extract from “Three Men in a Boat”, Jerome K Jerome (1889)

"… because the data is ***remote*** (and often ***replicated under the covers***), there's typically a ***noticeable latency cost*** involved in using it rather than local storage [...] the ideal would be to use local disk for each replica for the sake of lower latency."

Low throughput

Latency overheads

"EBS has got a lot better … [At Netflix] we [still] ***don't quite trust it*** for Kafka workloads […] going for ***instance type***"

Shared ➔ Crosstalk

"… unless you want to add ***more complexity*** for your operations team, choose […] direct-attached storage …."

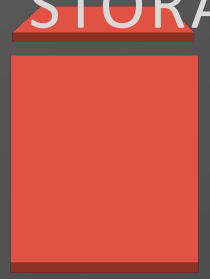
	Standard persistent disks	SSD persistent disks	Local SSDs	Cloud Storage buckets
Storage type	Efficient and reliable block storage	Fast and reliable block storage	High-performance local block storage	Affordable object storage
Price per GB/month	$0.04	$0.17	$0.218	$0.007 - $0.026
Maximum space per instance	64 TB	64 TB	3 TB	Almost infinite
Scope of access	Zone	Zone	Instance	Global
Data redundancy	Yes	Yes	No	Yes

	Standard persistent disks	SSD persistent disks	Local SSD (SCSI)	Local SSD (NVMe)
Maximum sustained IOPS				
Read IOPS per GB	0.75	30	266.7	453.3
Write IOPS per GB	1.5	30	186.7	240
Read IOPS per instance	3,000	40,000	400,000	680,000
Write IOPS per instance	15,000	30,000	280,000	360,000

(Roz Chast, The New Yorker, April 2001)

"[with Local SSDs] You cannot *stop and restart* an instance … only suitable for *temporary storage* …"

DECIBEL STORAGE GOALS



VS.

Managed Storage

Redundancy and high availability

Transparent device aggregation

Cluster-accessibility

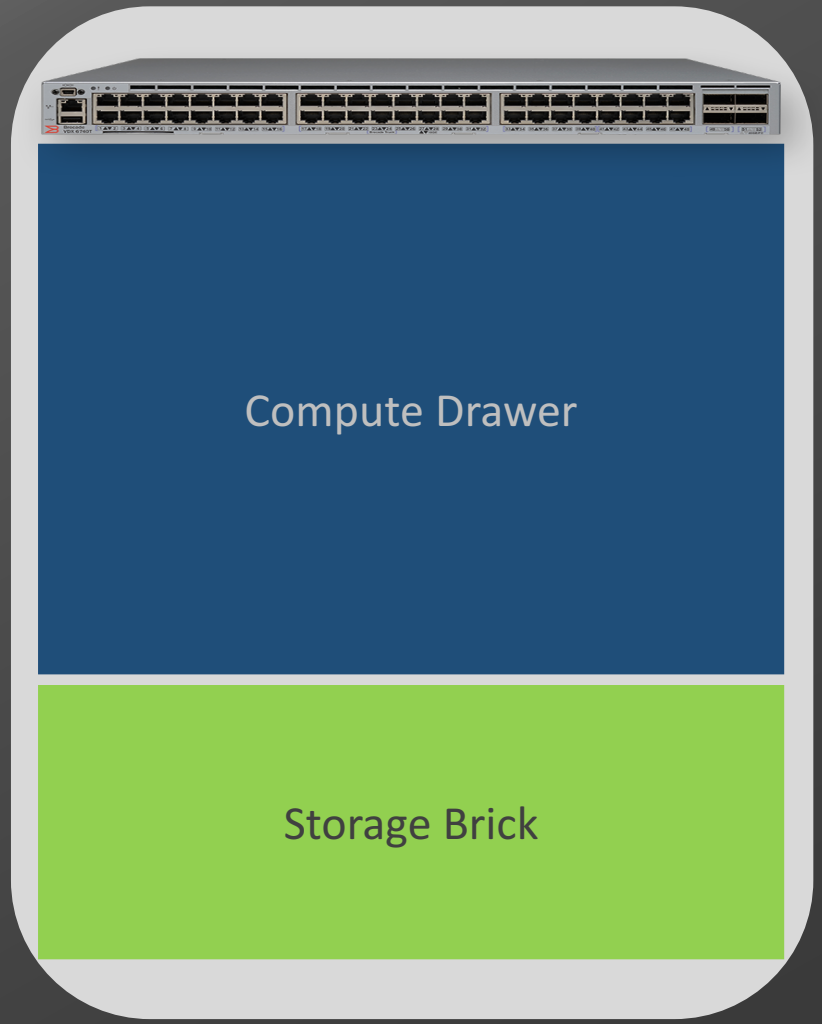
Decoupling and migration

Local Storage

Throughput and latency

No crosstalk

Tight host coupling



(Adapted from Intel Rack Scale Architecture)

Make all local storage cluster-accessible!

Decibel is a storage service that remotely serves “local” storage

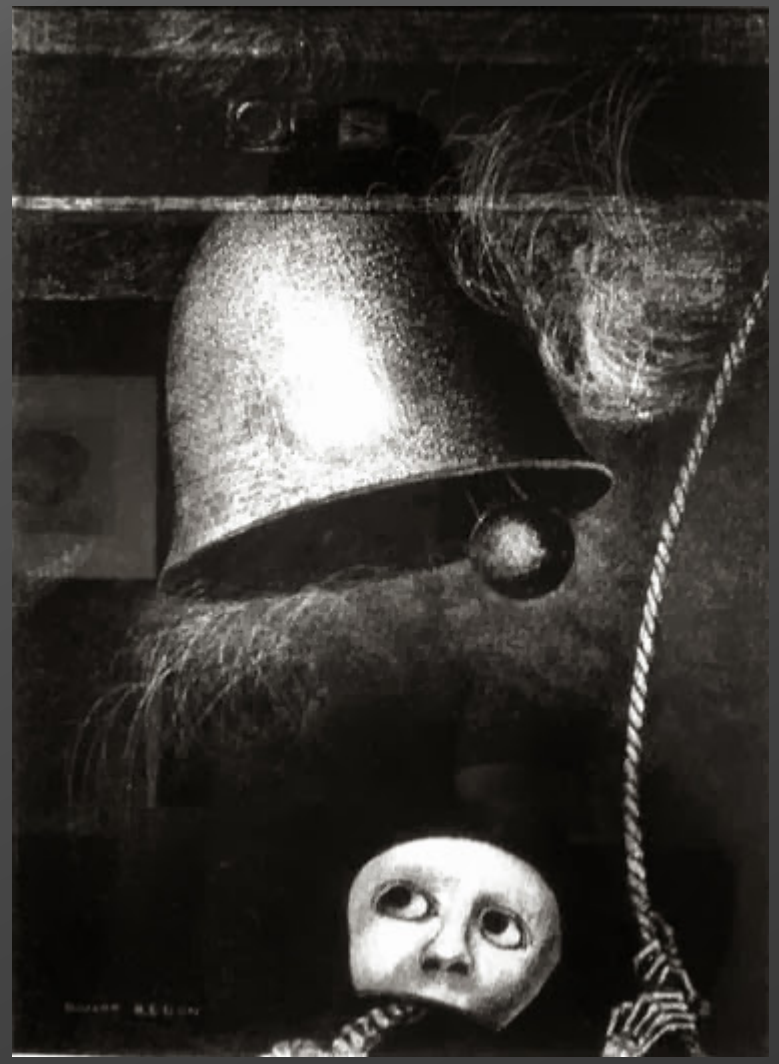
Challenges

Workloa

Virtualiz

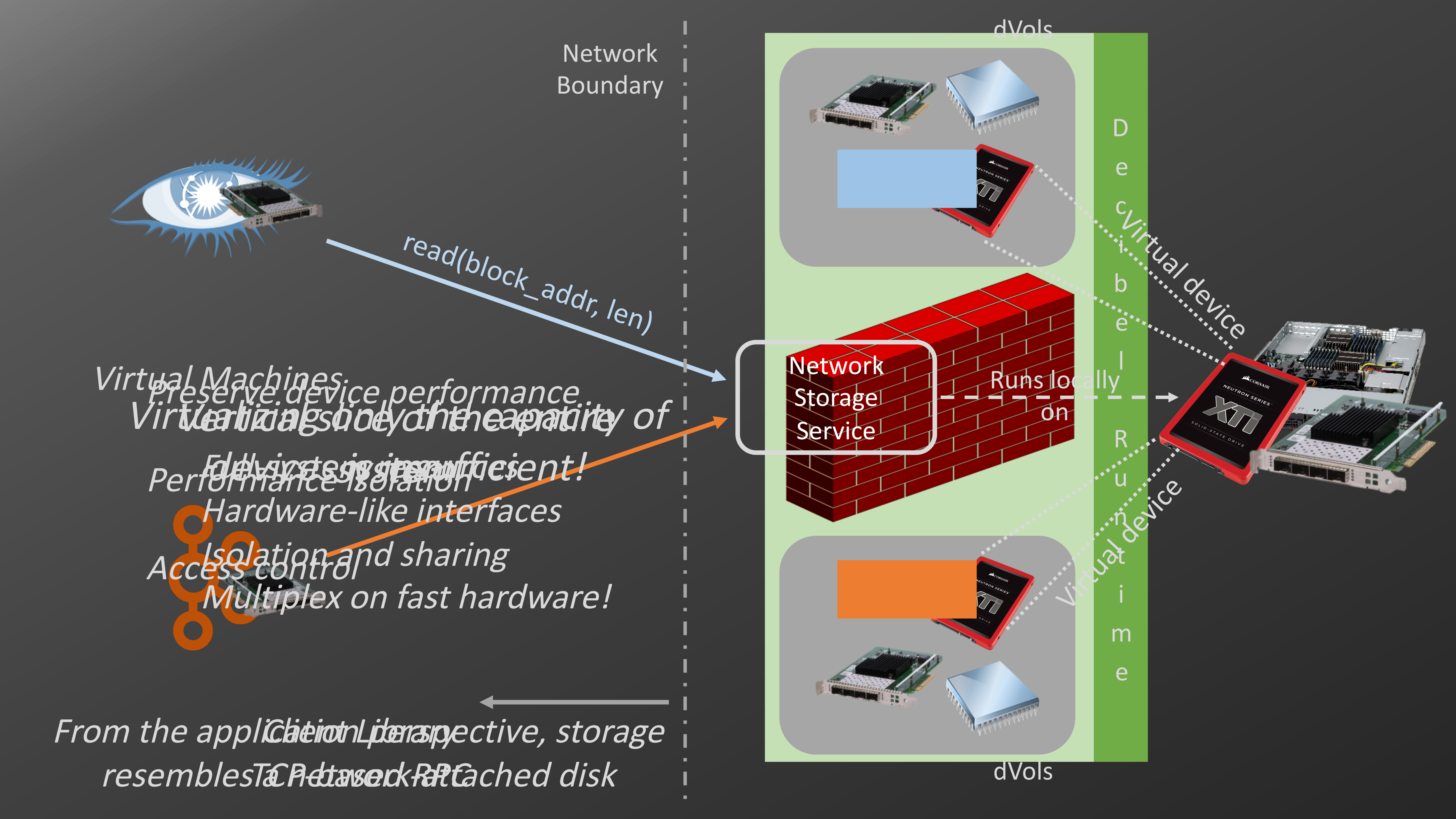
Maintain device performance

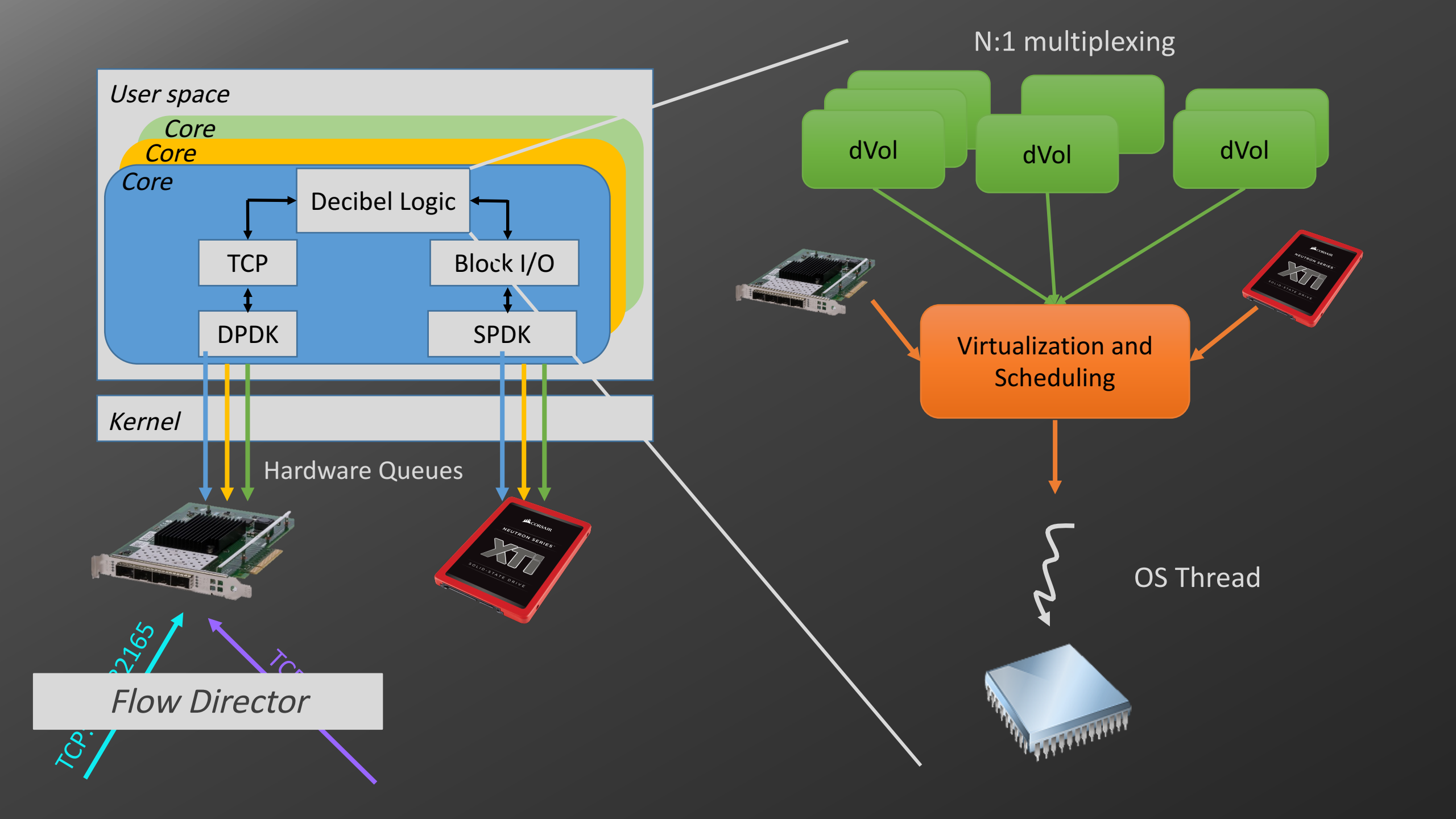
Virtualized Storage Abstraction + Runtime

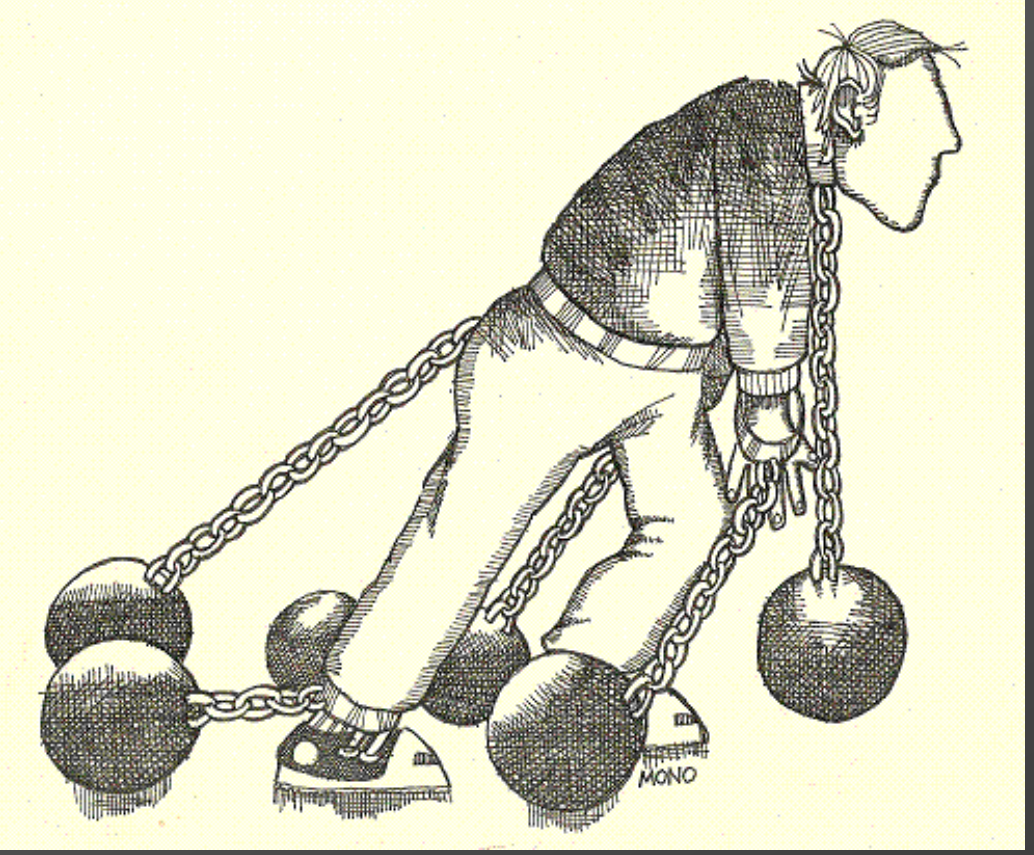


“They are neither man nor woman
They are neither brute nor human
They are Ghouls”

Extract from “The Bells”, Edgar Allan Poe (1849)


Network Boundary
read(block_addr, len)
Virtual Machines
Preserve device performance
Hardware-like interfaces
Isolation and sharing
Multiplex on fast hardware!
Performance isolation
Access control
From the application perspective, storage
resembles a network-attached disk
Client Library
TCP-based RPC
dVols
Network Storage Service
Runs locally on
Virtual device
Virtual device
Decibel Runtime
dVols


N:1 multiplexing
User space
Core
Core
Core
Decibel Logic
TCP
Block I/O
DPDK
SPDK
Kernel
Hardware Queues
dVol
dVol
dVol
Virtualization and Scheduling
OS Thread
Flow Director



(PD Lankovsky)

dVols are bound to a single core and a single device

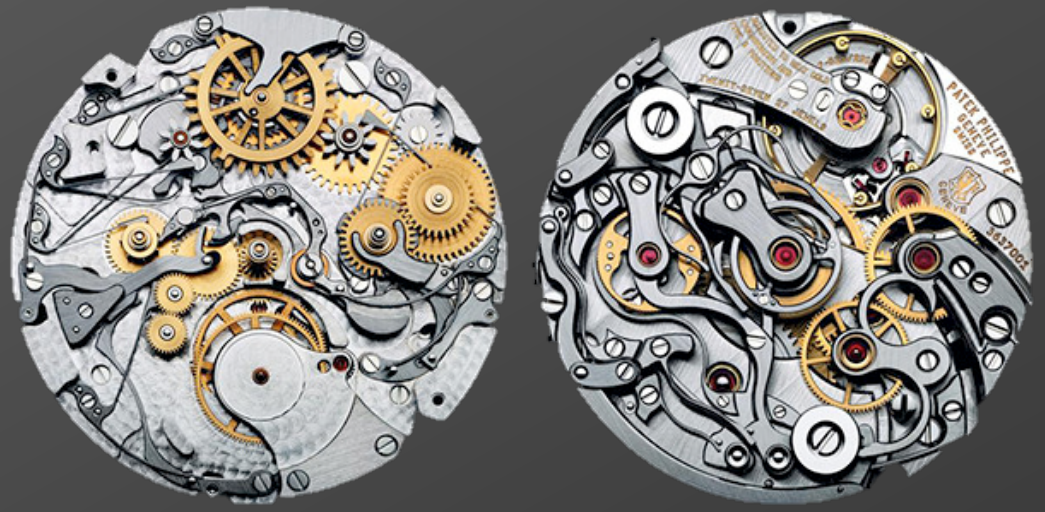
Limits mimic local SSDs

Decouple placement from scheduling

(Mirador, FAST 2017)

Capacity Migrations

Performance Migrations



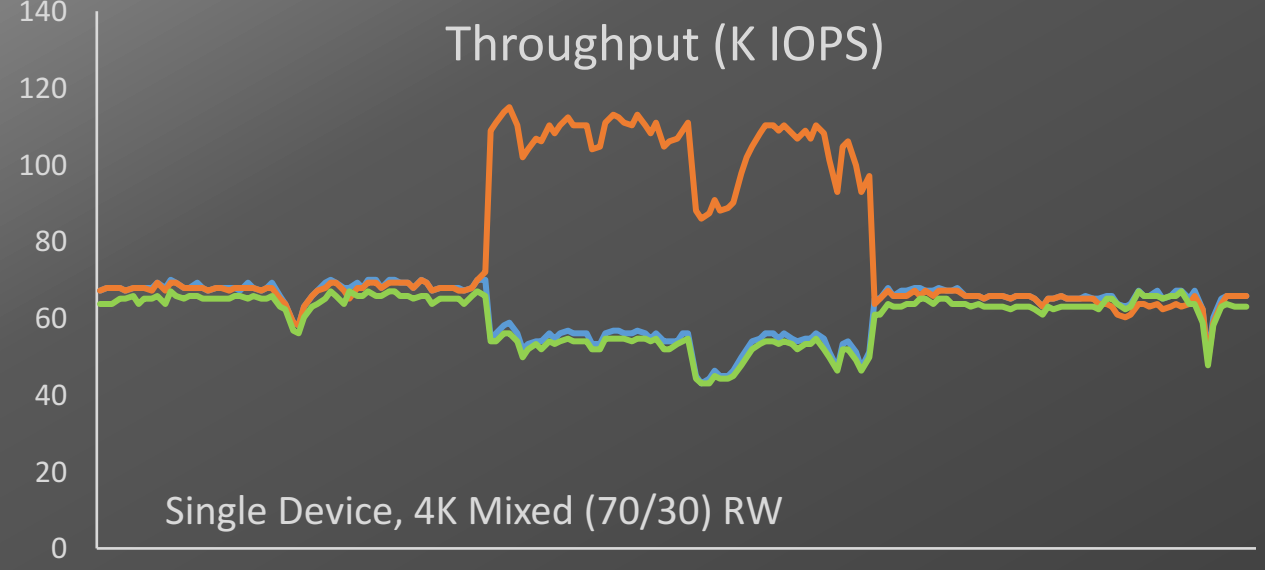
(Guido Mocafico, Movement)

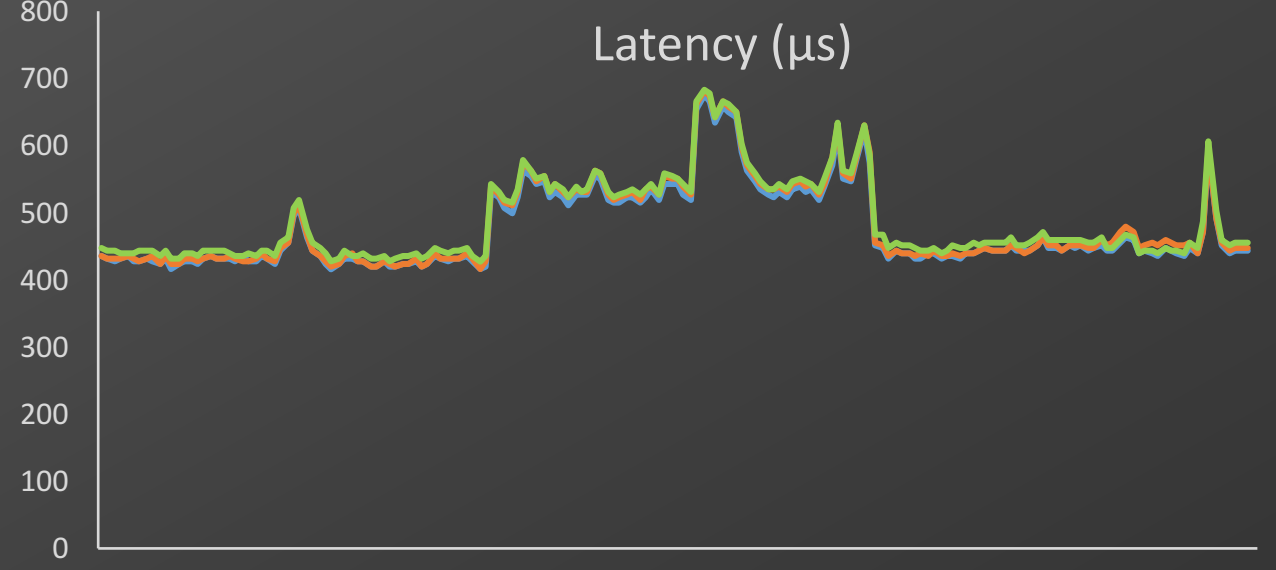
Device Partitioning

- Contention and fragmentation
- Per-core caches (Hoard/`tcmalloc`)

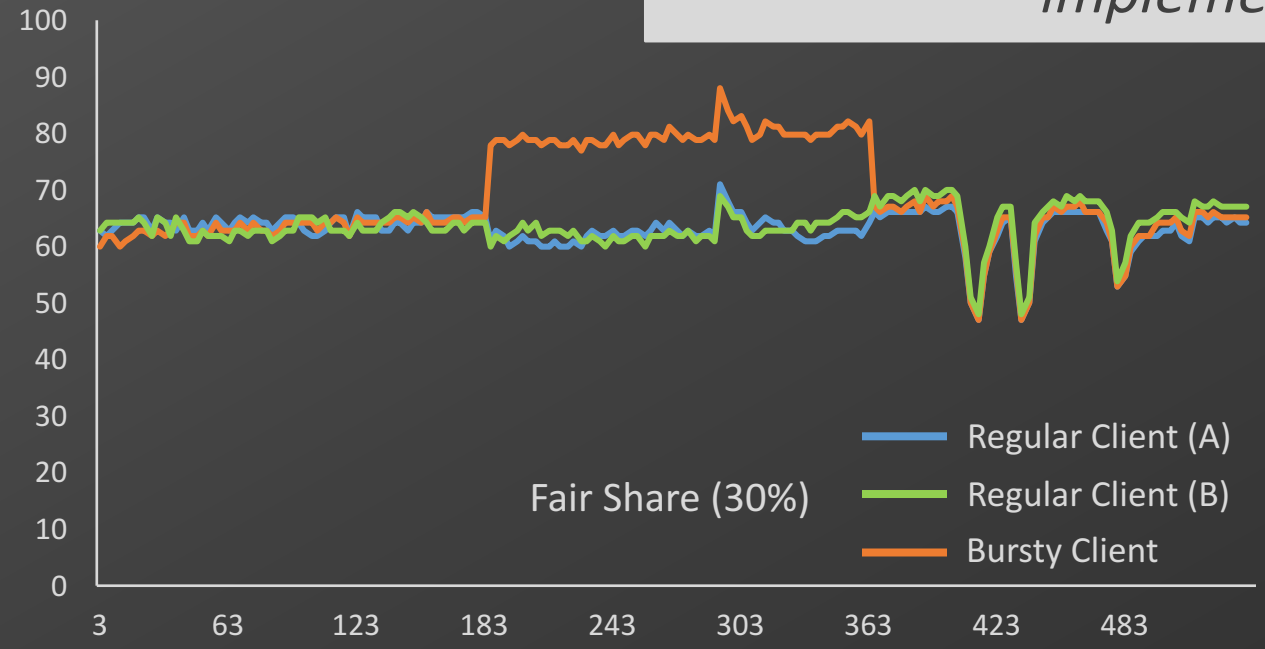
Scheduling

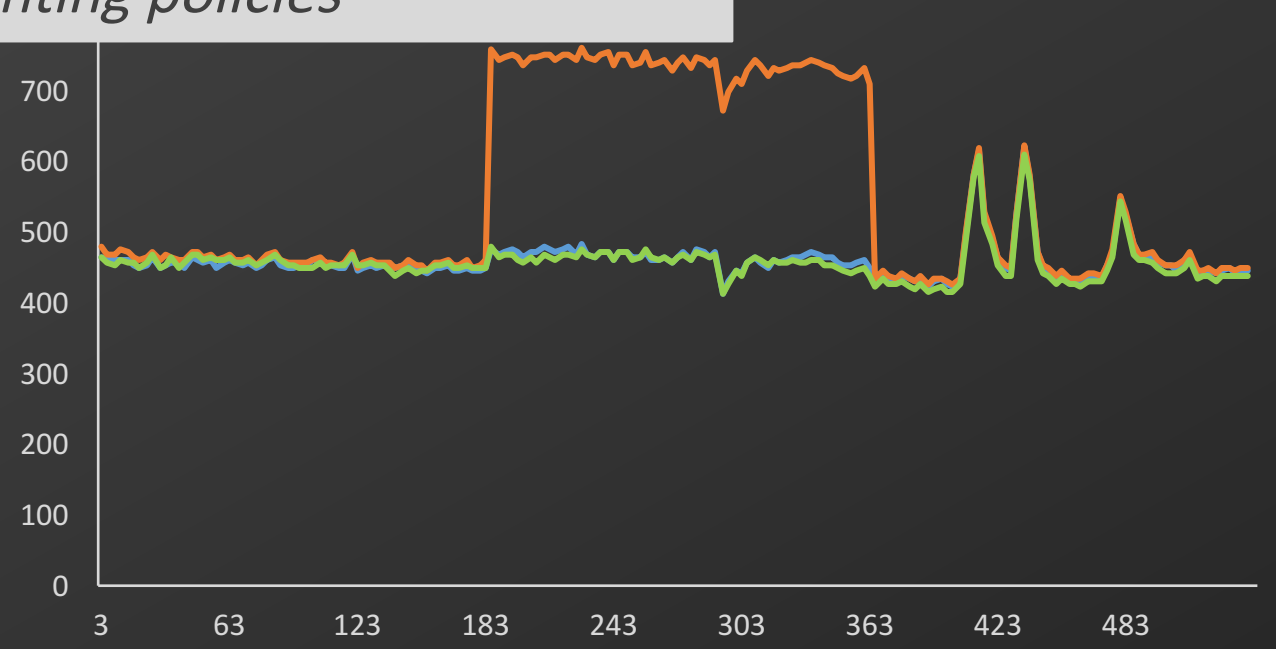
- Device throughput and latency
- dVol isolation





dVols provide hooks for conveniently implementing policies





Decibel (DRR)

THE DECIBEL RUNTIME

Resource Management

Scheduling

Address Virtualization

Access Control

Atomicity and Data Integrity

Decibel vs. Local Devices

Performance

Single-core Performance

THE SETUP

1U Client

1U Server
2 x Fortville 40GbE

4 x P3700
Clients/Core : 2

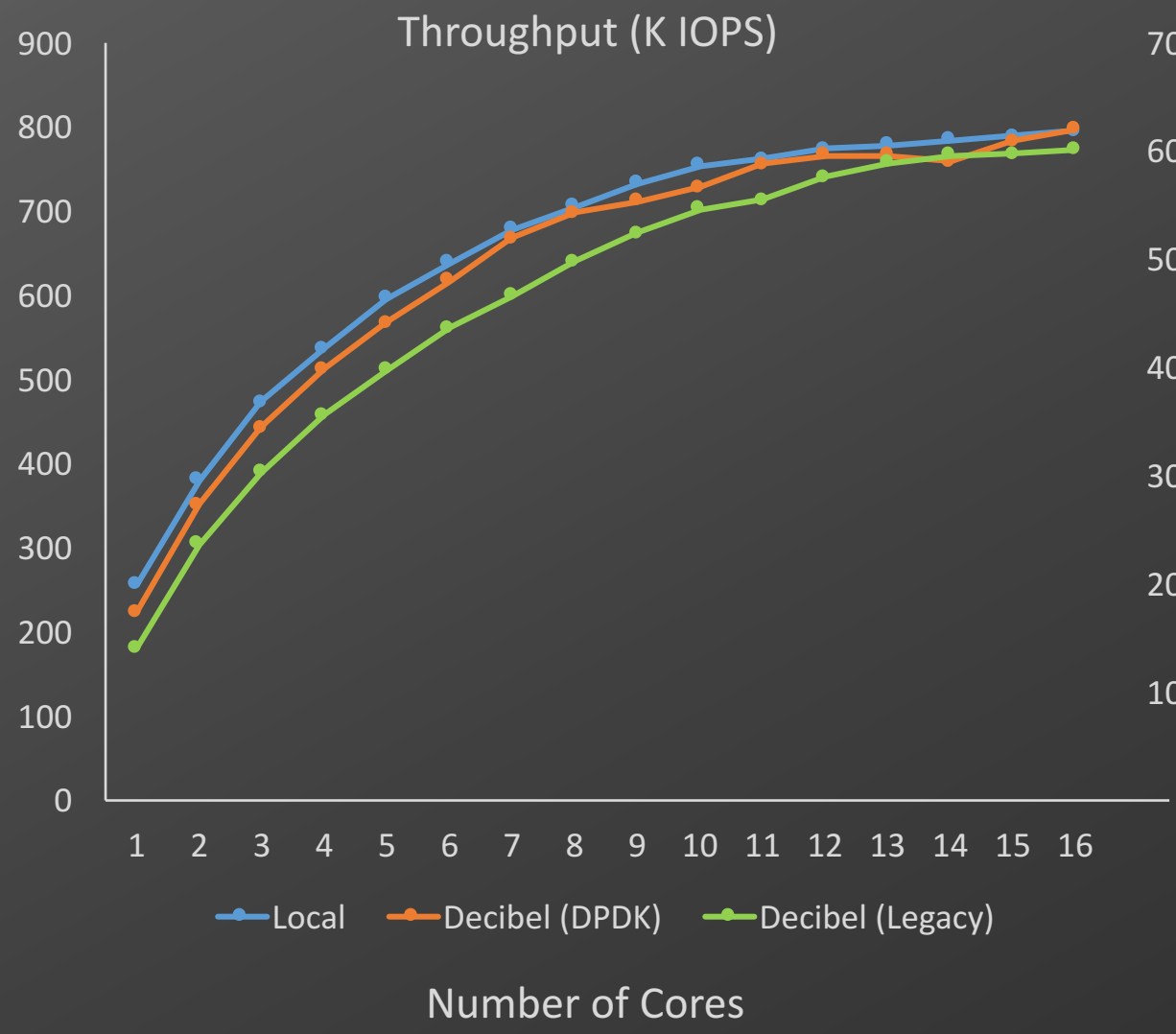
2 x Fortville 40GbE
QD/client : 16

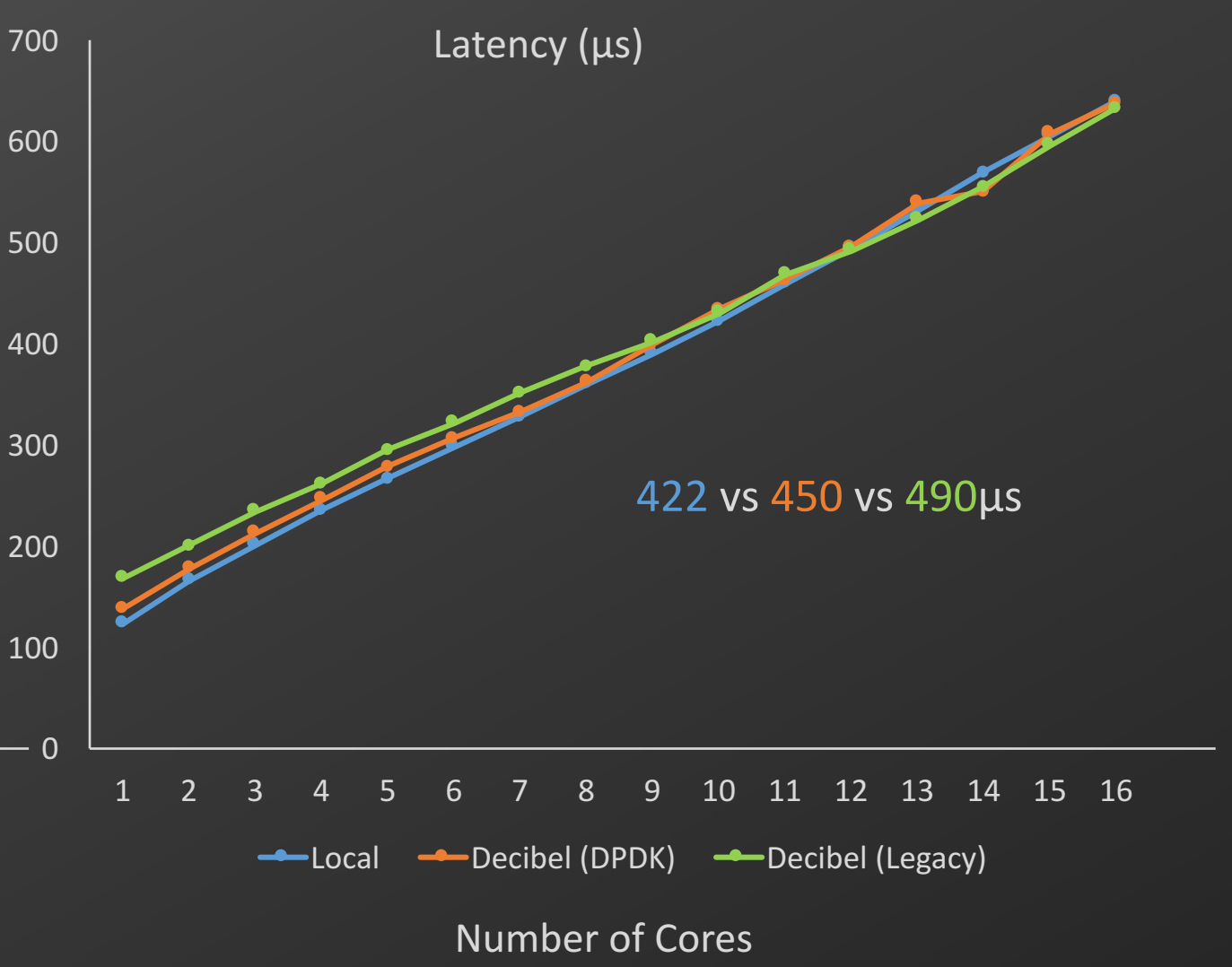
4K random requests

1.8 M IOPS (60 Gbps)

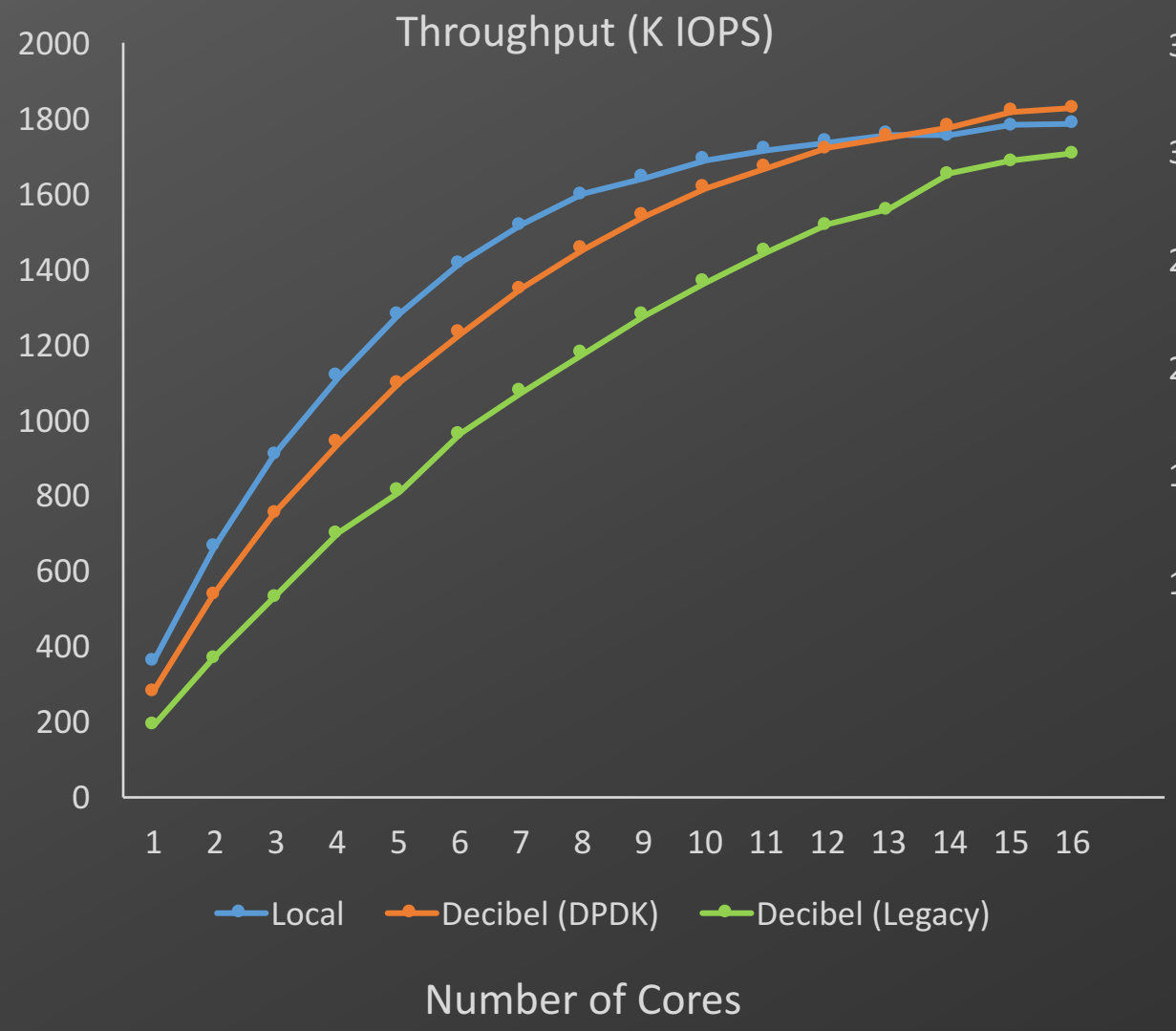
Independent remote dVol

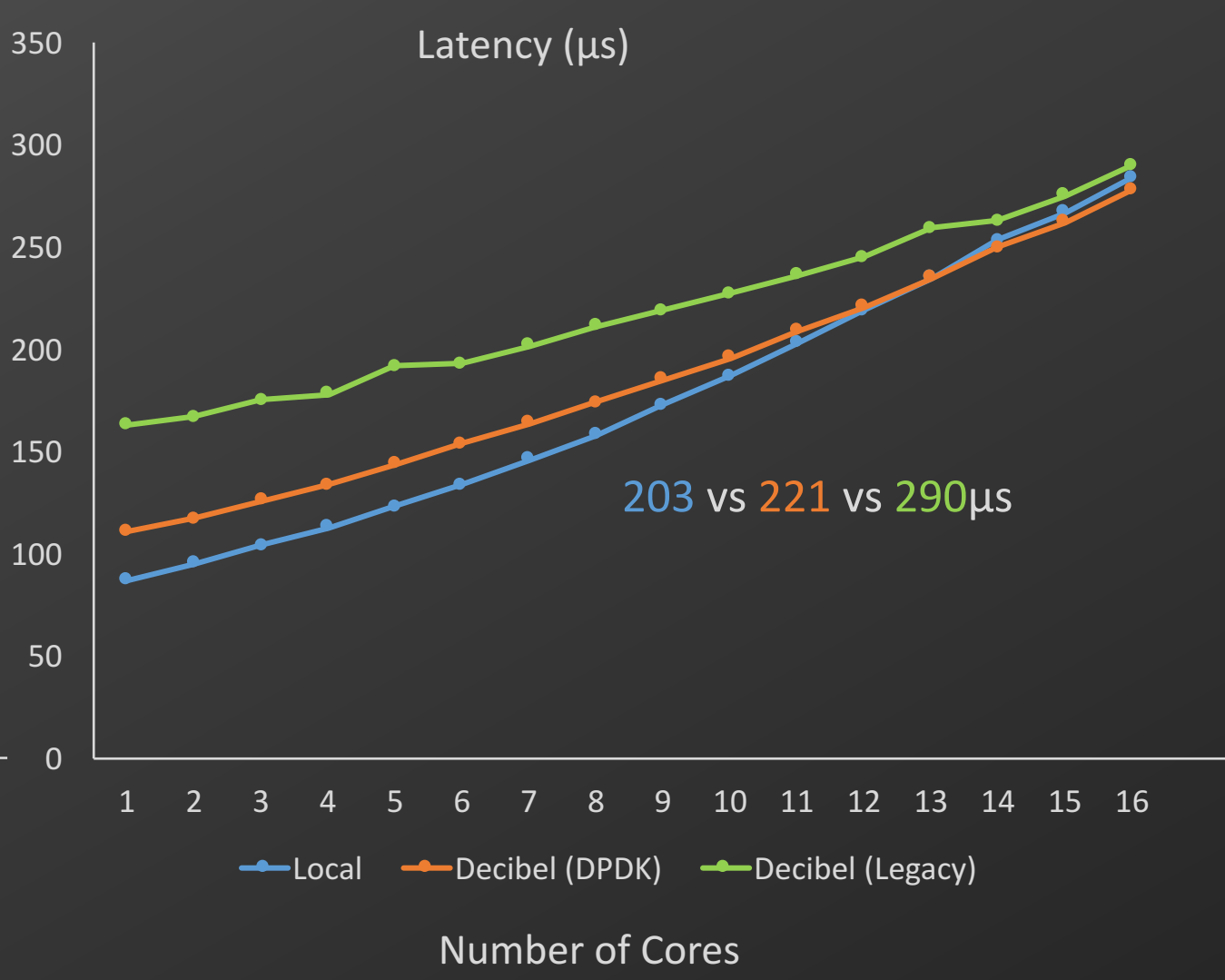
Performance (70/30 Mixed Workload)



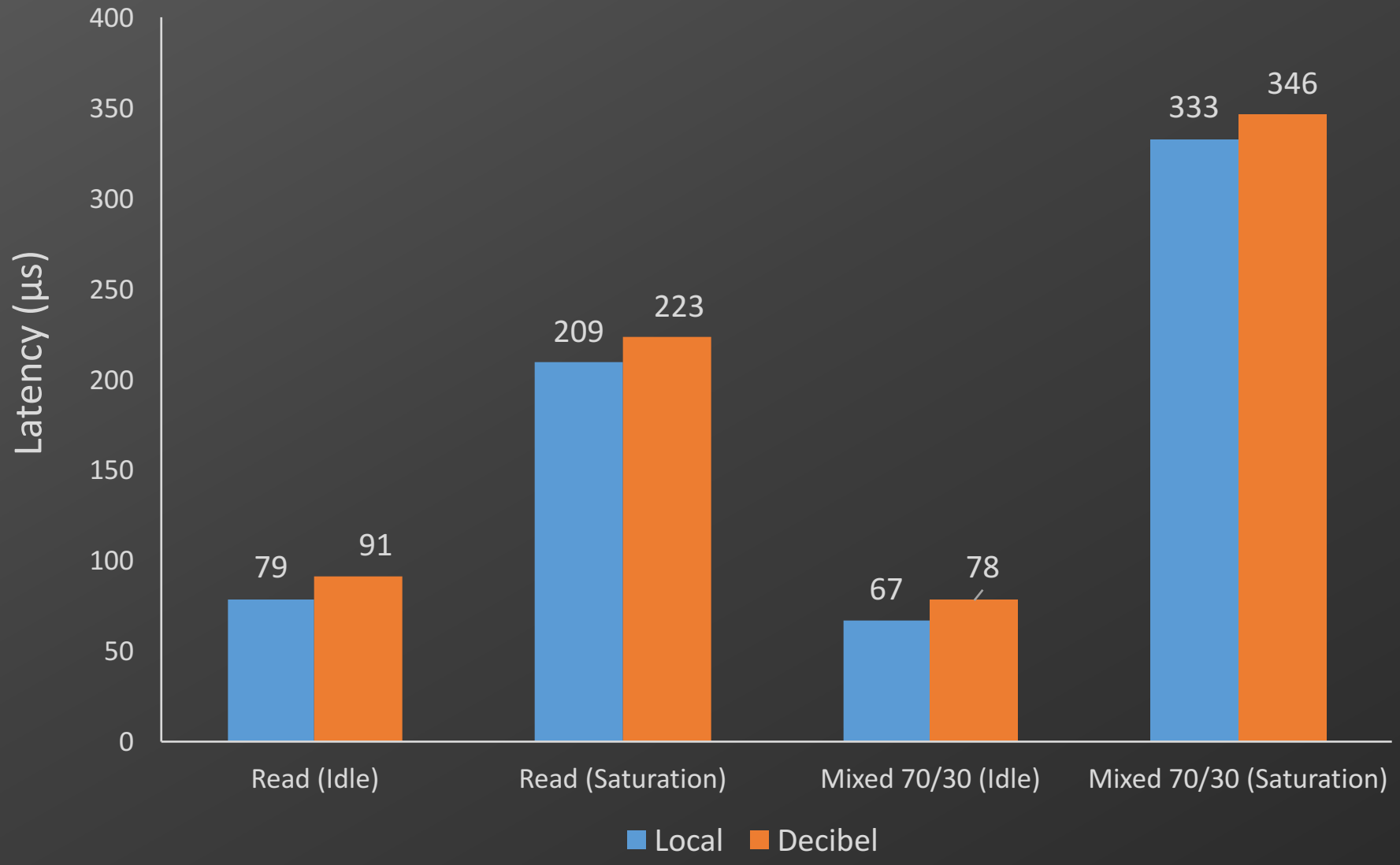


Performance (All Reads Workload)





SINGLE CORE PERFORMANCE



Local storage as a service

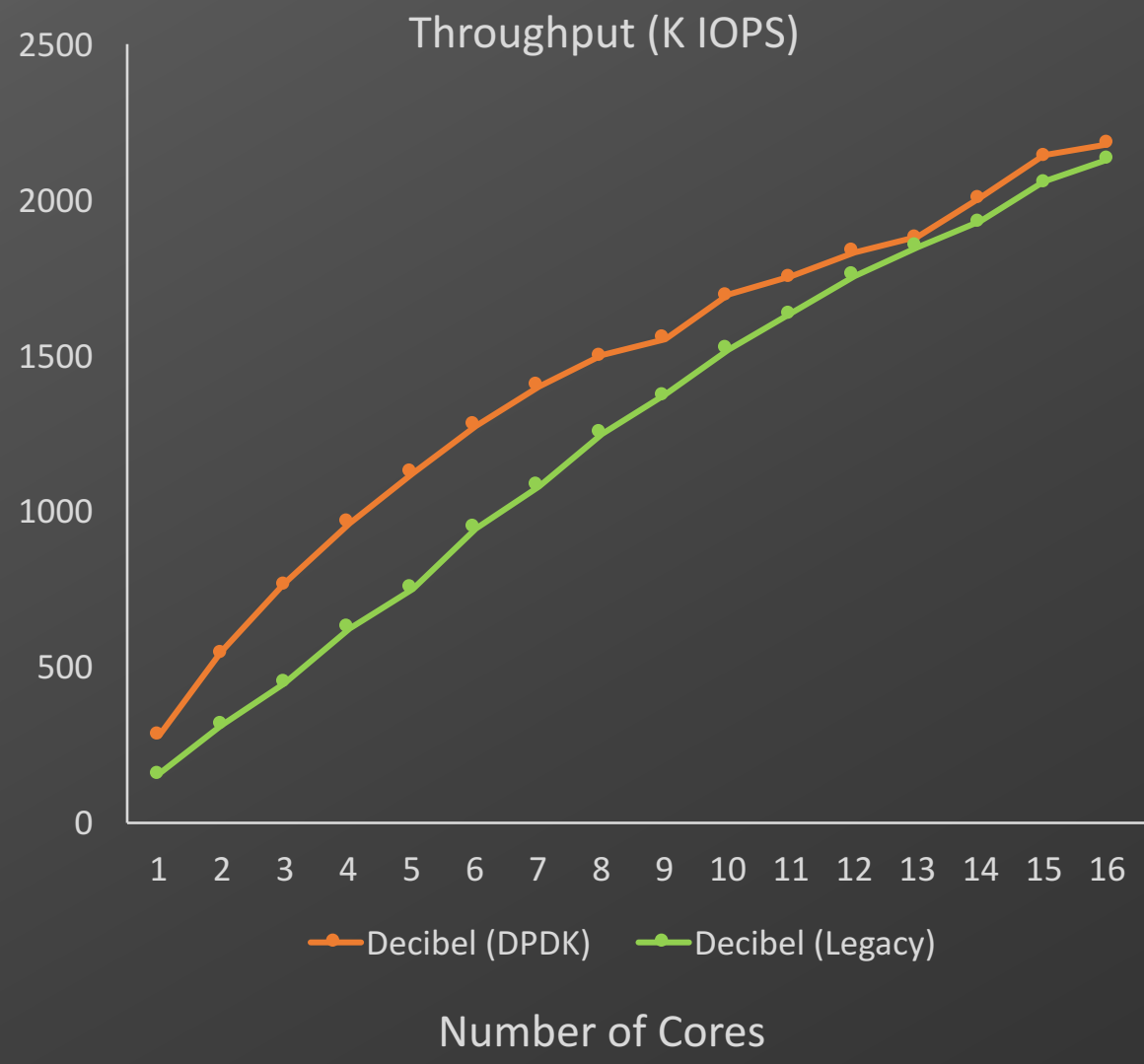
Encapsulate full-system resources

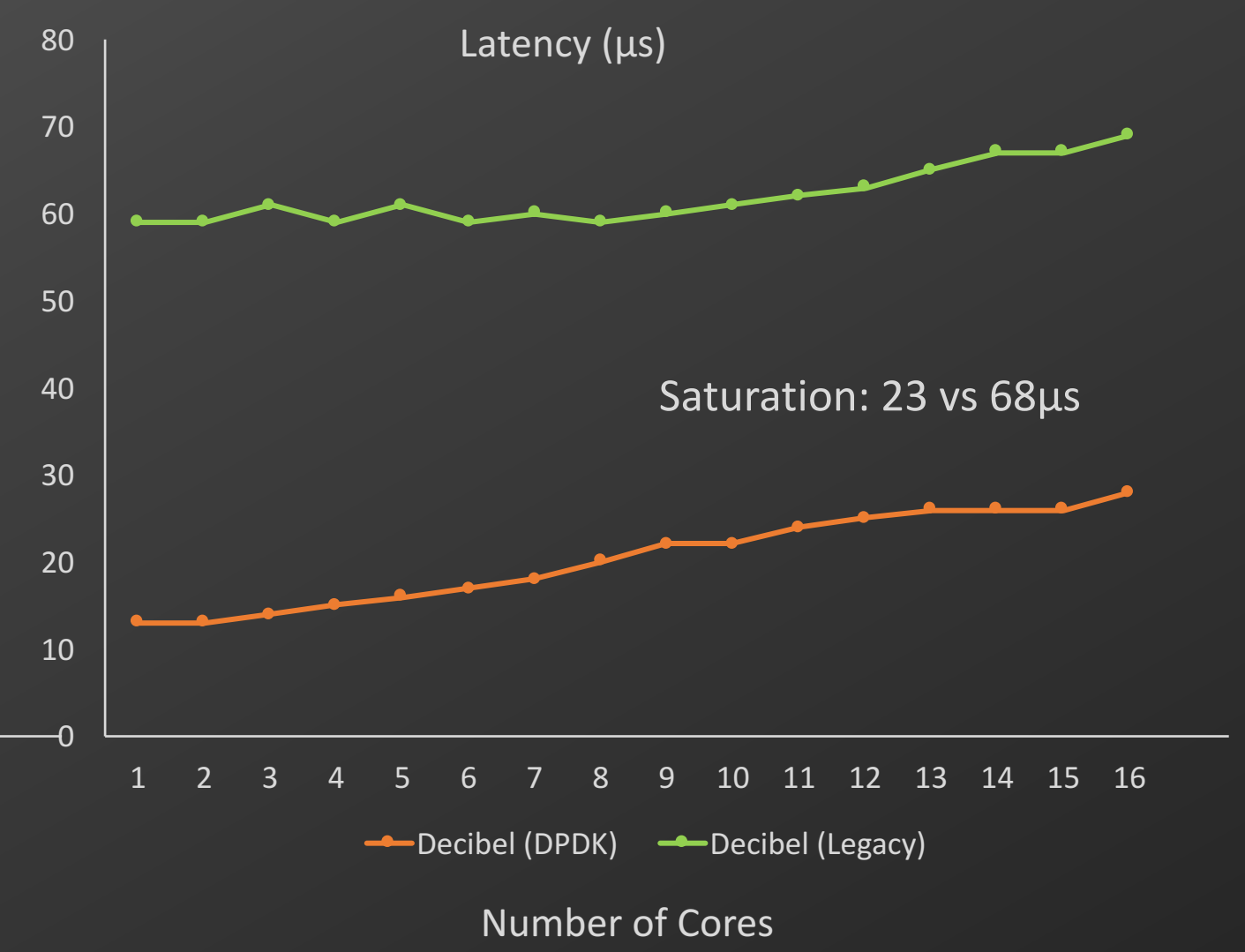
20-30µs overhead to local devices

Isolation on shared storage

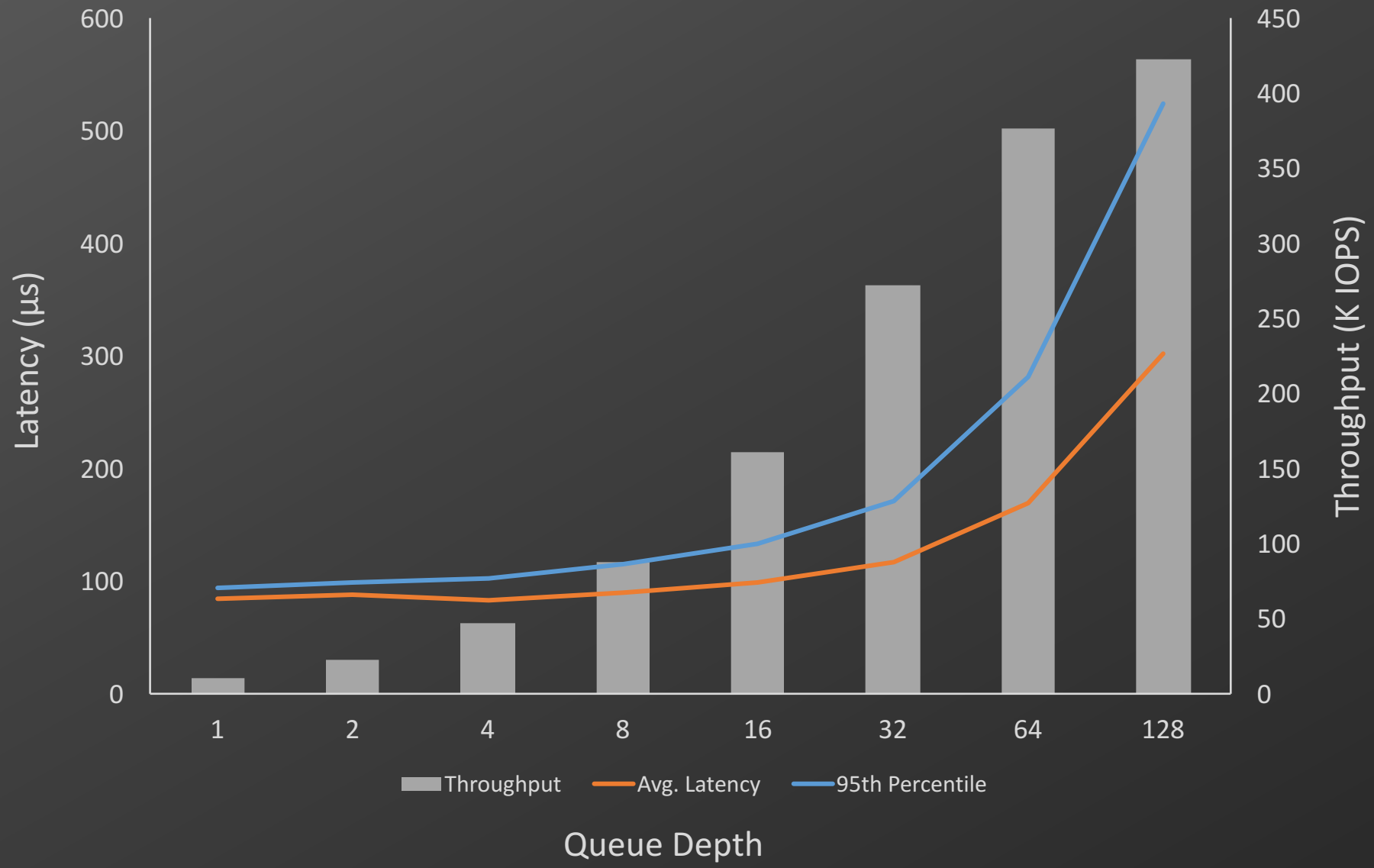
Backup Slides

Performance (DRAM Backed Storage)





LATENCY CONSISTENCY (READS)



LATENCY CONSISTENCY (WRITES)