SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Ceph: The Linux of Storage Today

Live Webinar

March 26, 2024

10:00 am PT / 1:00 pm ET

Today's Presenters

Erin Farr
Vice Chair SNIA Cloud Storage Technologies Initiative
Storage CTO Office, IBM

Tushar Gohad
Sr. Principal Engineer
Storage Software Architecture
Intel

Vincent Hsu
VP, IBM Fellow, and CTO for Storage and Software Defined Infrastructure
IBM

The SNIA Community

200
Corporations, universities, startups, and individuals

2,500
Active contributing members

50,000
Worldwide IT end users and professionals

SNIA™ CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements Incorporate them into standards and programs

Collaborate with other industry associations

SNIA Legal Notice

- The material contained in this presentation is copyrighted by SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

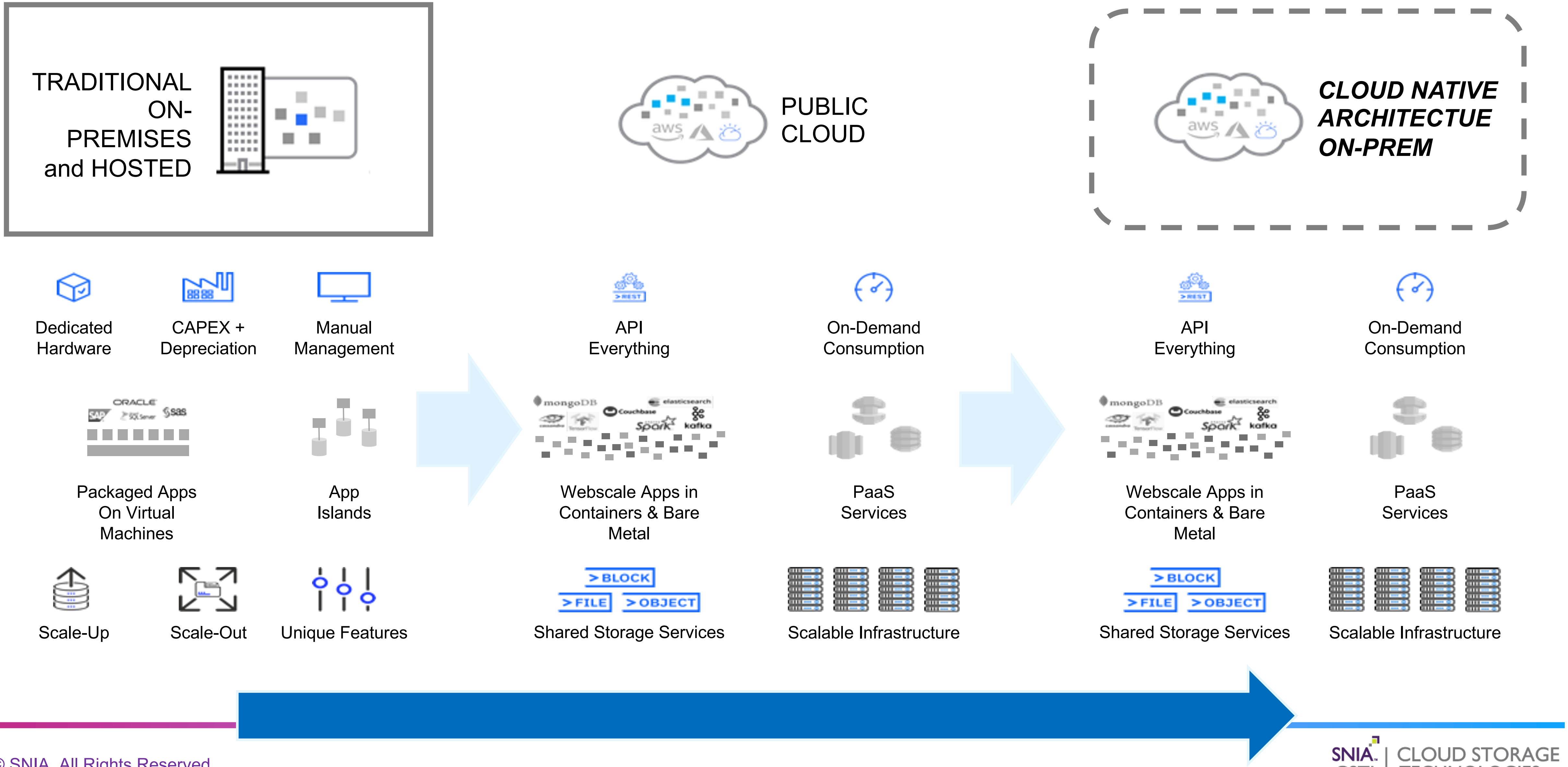
- The evolution of the data center
- Ceph as the primary storage
- Consumability
- Resiliency and Security
- Crimson: The Next Generation Ceph OSD
- Storage for AI

Evolution of the Data Center

Vincent Hsu

IBM

Introducing: “Cloud Native Architecture On-prem”



TRADITIONAL ON-PREMISES and HOSTED

Dedicated Hardware

CAPEX + Depreciation

Manual Management

Packaged Apps On Virtual Machines

App Islands

Scale-Up

Scale-Out

Unique Features

PUBLIC CLOUD

API Everything

On-Demand Consumption

Webscale Apps in Containers & Bare Metal

PaaS Services

Shared Storage Services

Scalable Infrastructure

CLOUD NATIVE ARCHITECTUE ON-PREM

API Everything

On-Demand Consumption

Webscale Apps in Containers & Bare Metal

PaaS Services

Shared Storage Services

Scalable Infrastructure

Ceph: unified SDS for primary storage

RBD NVMe over TCP for VMware

Performance
Low latency of NVMe while maintaining the flexibility of TCP

Simplicity
A new management layer in Ceph simplifies the configuration of targets across multiple cluster nodes

Flexibility
Pure user-space implementation enables multiple topologies and dynamic scale

RGW Best in class Object

Scalable & Durable
Exabyte scale at maximum throughput, all-flash configurations, multi-site replication and backup

Standard
Best in class S3 fidelity for integration with modern applications

Simple
Web based management, Autonomic balancing and self-healing, Ceph ready-node architecture

CephFS File storage

CephFS directly leverages the scalability, parallelism, performance and reliability of Ceph's core data engine: RADOSCompatible & Flexible File Storage for Legacy UNIX:

- Use NFS v3 and v4 to connect legacy servers

Unified Storage experience for File & Object:

- Ingest / Export Files via NFS into/from the Ceph Object Store and use S3 for compute

Use native & high performance CephFS for Linux, OpenShift, OpenStack

Reliable Autonomic Distributed Object Store (RADOS)
RADOS Block Devices (RBDs)
RADOS Gateway (RGW)

Ceph: enterprise ready SDS for cloud scale primary storage

Consumability

1. Easy installation setup
2. Centralized management
3. Easy upgrade process
4. Easy Scalability
5. Easy start process

Resiliency & Security

1. STS: Security Token Services
2. IAM roles bucket services
3. S3 table with data encryption

Performance

1. Crimsom
2. Seastore

Consumability

Easy Installation Setup

Ceph initial cluster setup

- Start with a minimum of 4 industry-standard x86-servers running Ceph and easily scale out to actual business needs.

Ceph software internals

- Ceph software internally runs Linux containers, removing any needs for specific dependencies.

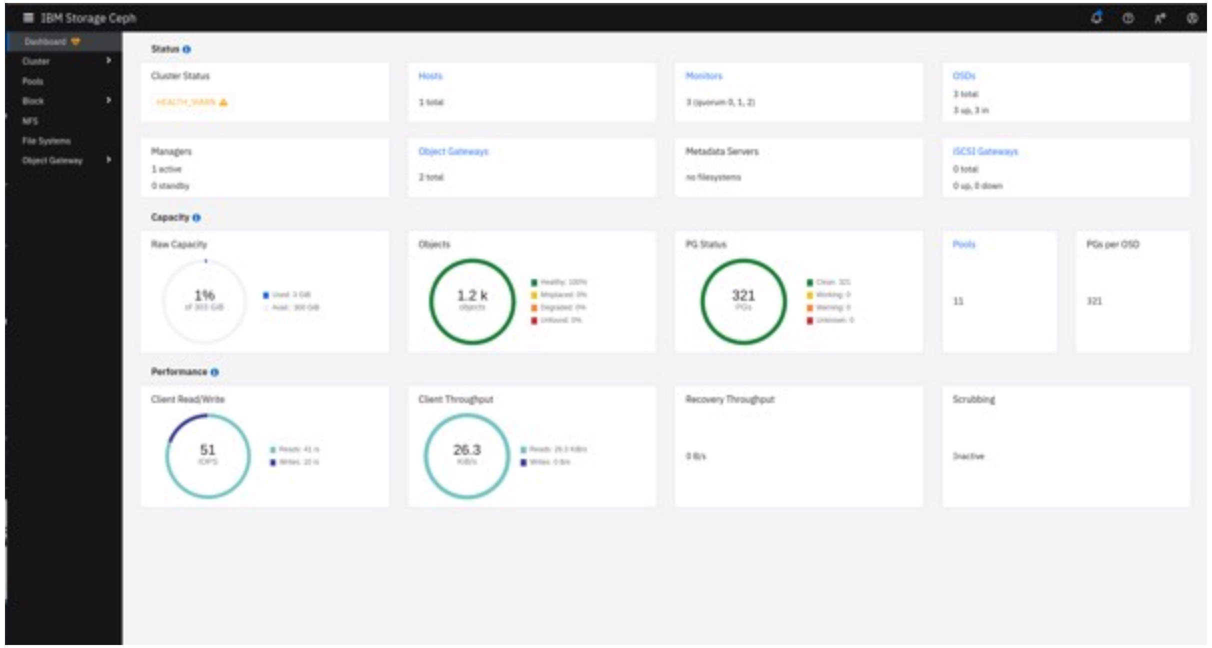
More flexible, faster and easier to deploy and maintain, compared to conventional package-based software deployment.

Ceph single command installation setup

- An Ceph cluster can literally be installed by running one single command.

With this, executing the installation process has become as simple as pressing one button.

Easy Centralized Management



Status

Cluster Status: HEALTH_WARN

Hosts: 1 total

Monitors: 3 (quorum 0, 1, 2)

OSDs: 3 total, 3 up, 3 in

Managers: 1 active, 0 standby

Object Gateways: 2 total

Metadata Servers: no filesystems

iSCSI Gateways: 0 total, 0 up, 0 down

Capacity

Raw Capacity: 1% of 303 GiB (Used: 3 GiB, Avail.: 300 GiB)

Objects: 1.2 k objects (Healthy: 100%, Misplaced: 0%, Degraded: 0%, Unfound: 0%)

PG Status: 321 PGs (Clean: 321, Working: 0, Warning: 0, Unknown: 0)

Pools: 11

PGs per OSD: 321

Performance

Client Read/Write: 51 IOPS (Reads: 41 /s, Writes: 10 /s)

Client Throughput: 26.3 KiB/s (Reads: 26.3 KiB/s, Writes: 0 B/s)

Recovery Throughput: 0 B/s

Scrubbing: Inactive

CLI-based administration tool cephadm

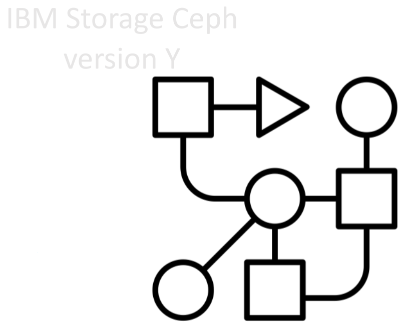
- Cephadm is a utility that deploys and manages an Ceph cluster.
- Cephadm is tightly integrated with both the command-line interface (CLI) and the Ceph dashboard through a web user interface.
- Clients can manage Ceph clusters from within either environment

Ceph user interface dashboard

- An intuitive user interface that allows for easy, straightforward navigation

  Delivers an easy point and click experience for common administrative tasks
- In example, managing storage capacity, configuring services, access for file, block and object, object buckets, users, and S3 access keys.

Easy Application Integration



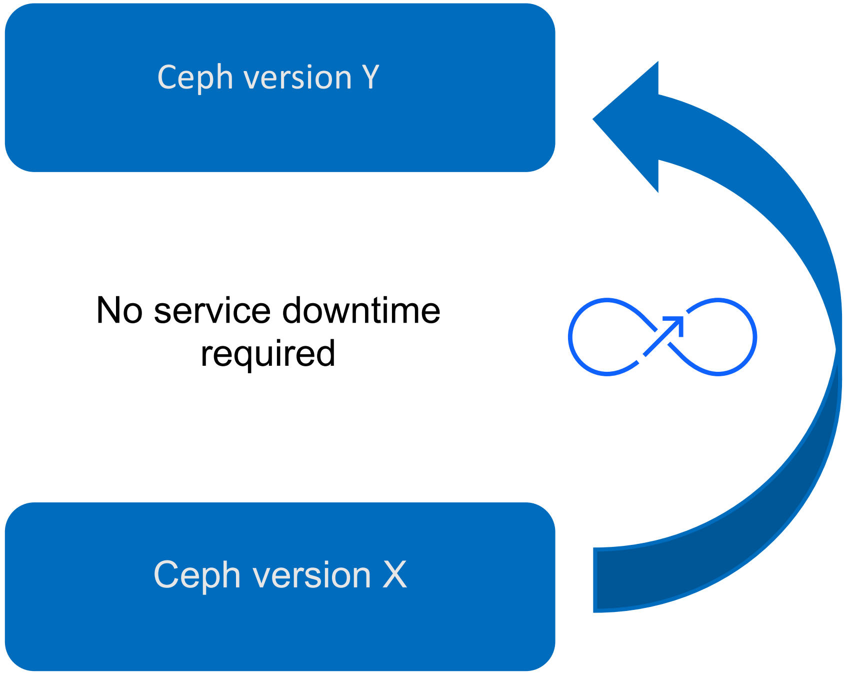
Ceph application integration

- Ceph offers File, Block and Object APIs and protocols to support a broad variety of applications and platforms.
- Ceph features a set of industry standard and consistent APIs for data and service consumer interactivity.

Ceph standardized APIs

- Application developers require assurance that they can rely on a storage solution that provides consistent and standardized APIs such as S3 for Object or POSIX for File.
- Ceph leads in the on-premise object storage market when it comes to AWS S3-fidelity and compatibility.

Easy Upgrade Process



Updates and upgrades without downtime

- Afterwards initial Ceph cluster deployment, the process to update and upgrade Ceph software is similar to upgrading a firmware-image of a legacy storage system.

  However, with this one major difference and advantage:

- No downtime required!

Node-by-node online update process

- Ceph clusters will remain online while the software containers are upgraded node-by-node.
- Instead of hours or days of installing packages no service downtime is required at all.
- Internal software container images update and execute quickly and smoothly.

Easy Scalability

Ceph easily meets growing demands

- As a distributed storage system, Ceph scales effortlessly, to meet growing data demands and business needs.
- New nodes or devices can be added to the cluster without having disruptions or service downtime.

Ceph scale-out architecture

- Ceph scalability works in two dimensions: capacity and throughput.
- This straightforward scalability feature enables organizations to adapt to evolving storage needs seamlessly.

Ceph scalable management

- Scalable management capabilities to manage ever growing amounts of unstructured data.
- Alongside scalable cluster capacity and throughput, management of the cluster resources can also scale accordingly.

Resiliency and Security

STS
Security Token Service

About STS

- Ceph provides Security Token Service (STS).
- STS enables clients to request temporary and limited-privilege credentials for users.

Implementation

- Implements Amazon AWS compatible STS APIs.
- Related to cross account access and web identity federation

Supported actions

- AssumeRole
- AssumeRole WithWebIdentity
- GetSessionToken

Summary

- Returns temporary and limited privilege credentials, based on Amazon AWS Security Token Service
- Allows for integration with enterprise IDP authentication providers.

IAM Roles
Bucket Policies

Bucket policies

- Use bucket policies to grant permission to other users to access your S3 buckets.

Identity Access Management (IAM) role policies

- During STS authentication users can request to assume a role and inherit all the S3 permissions configured for that role, by RGW administrator

Share data in a secure way

- Role-based Access Control (RBAC) auth policies
- Attribute-based Access Control (ABAC) policy-based access control for IAM

Summary

- Prevent sharing S3 long lived passwords administration per user.
- Abstract permission settings and capabilities with bucket policies and IAM role policies.

S3 Tables with Data Encryption

Application storage

- Applications can access their storage through the Ceph S3 object API.
- Analytics tables can also be stored by using S3.

Encryption options

- Implementation of cluster-wide, at-rest, or user-managed inline object encryption.
- Managed encryption keys are supported.

SSE-S3 Serverside encryption

- AWS SSE-S3 similar functionality for Ceph on-premises or hybrid use, with UI management options

SSE-S3 Functionality

- RGW automatically encrypts objects before storing to disk and decrypts when objects are retrieved.

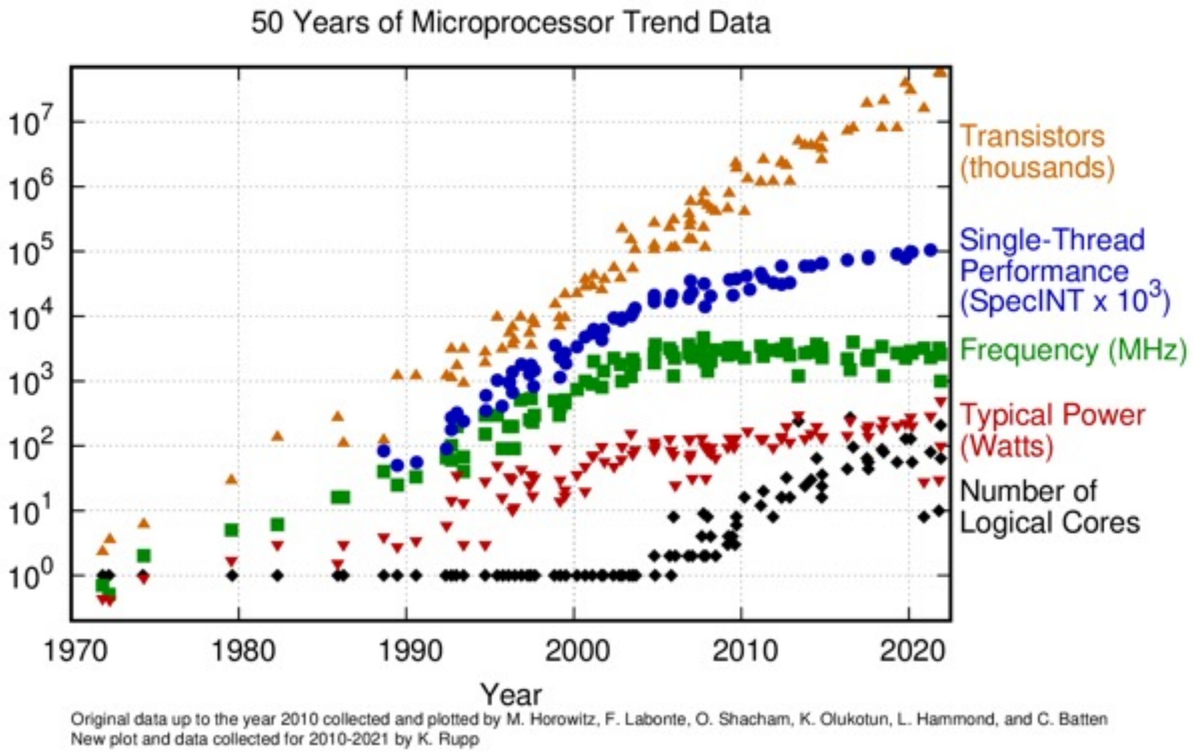
Crimson: The Next-Generation Ceph OSD

Tushar Gohad

Intel

Trends and Laws of Physics

- Core count has grown exponentially (1 -> 256)
- Networking and IO has gotten significantly faster
  - Disk seek time: 20ms -> 20us
  - Disk IOPS: 100 -> 1000000+
  - Network: 10Gbps -> 400+ Gbps
- Per-core performance plateaued

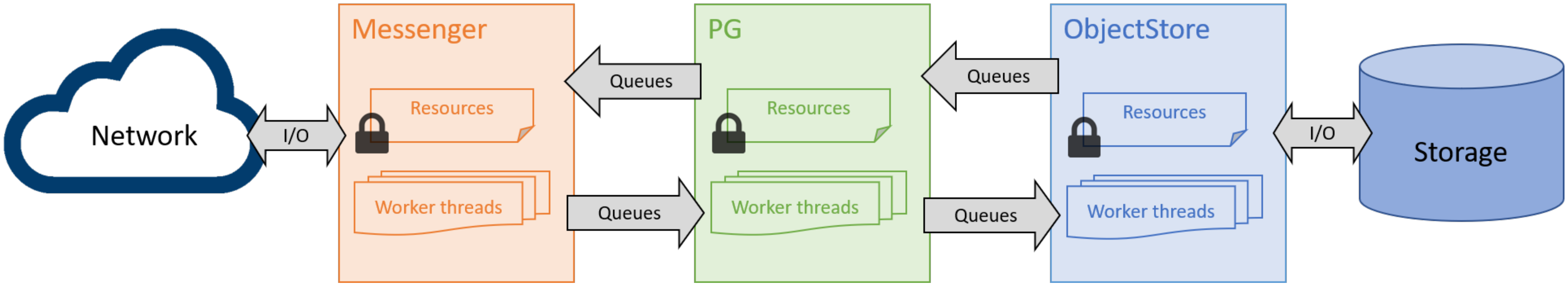


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp

https://github.com/karlrupp/microprocessor-trend-data

Classic Ceph OSD Architecture (Simplified)

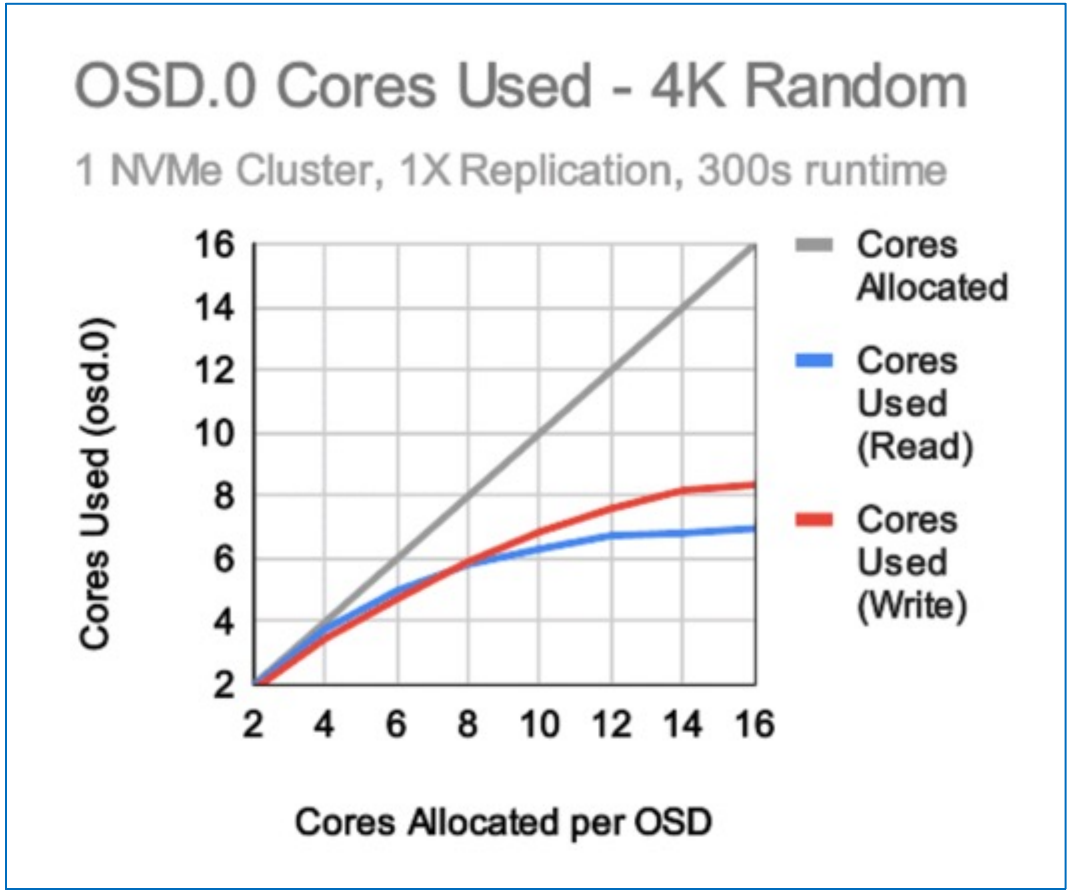
The Classic OSD

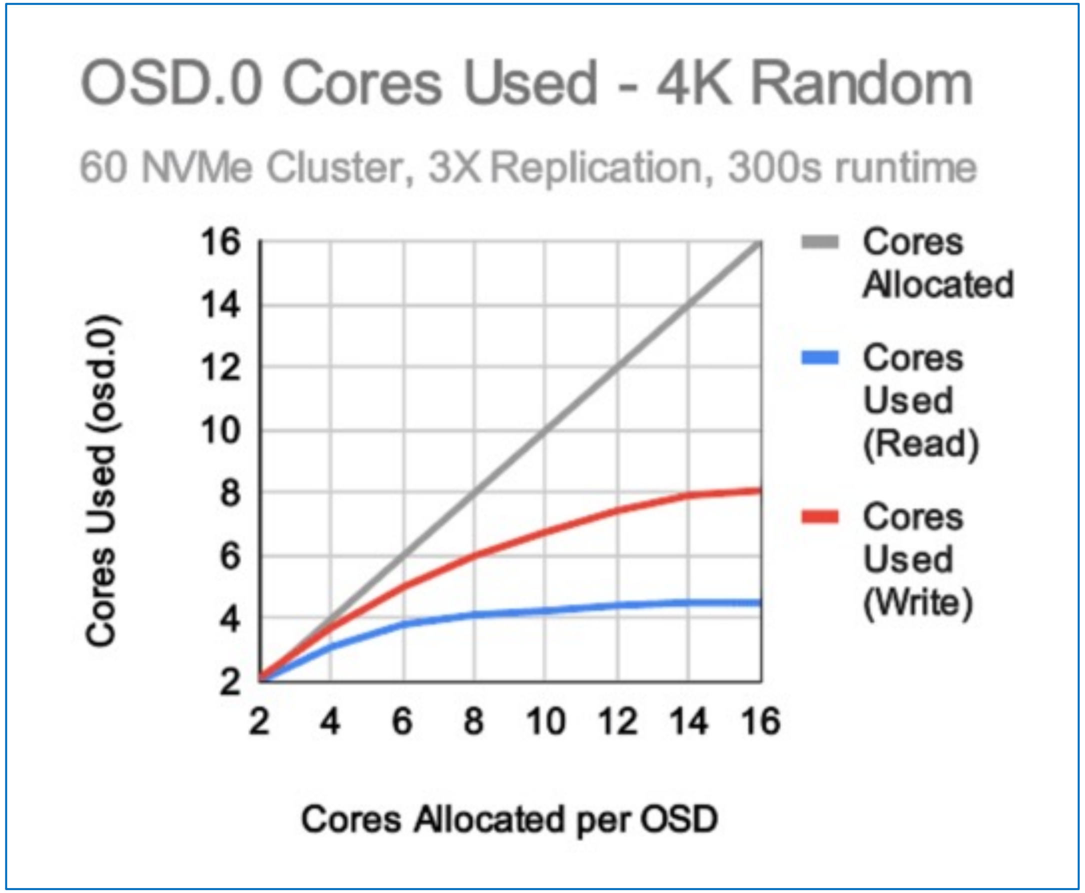


Thread pool with task queues, preemptive scheduling

Context switches, locking, shared resources

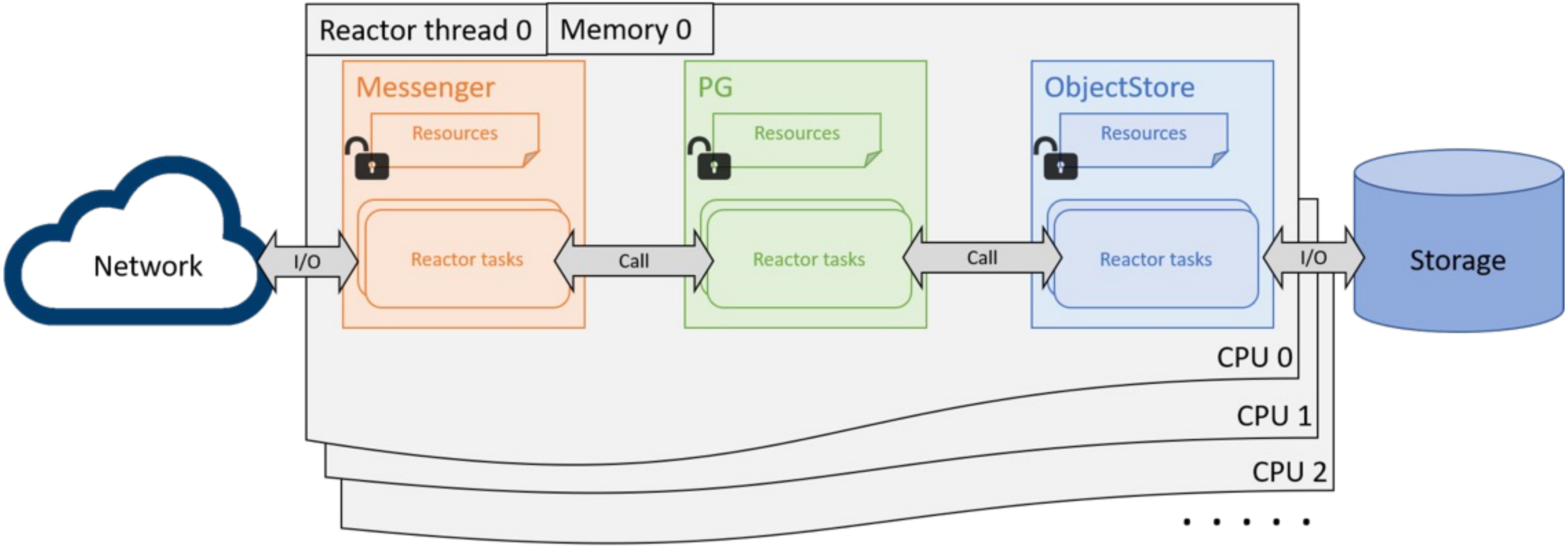
Classic OSD Core Scalability





Mark A Nelson, Ceph OSD CPU Scaling - Part 1, https://ceph.io/en/news/blog/2022/ceph-osd-cpu-scaling/

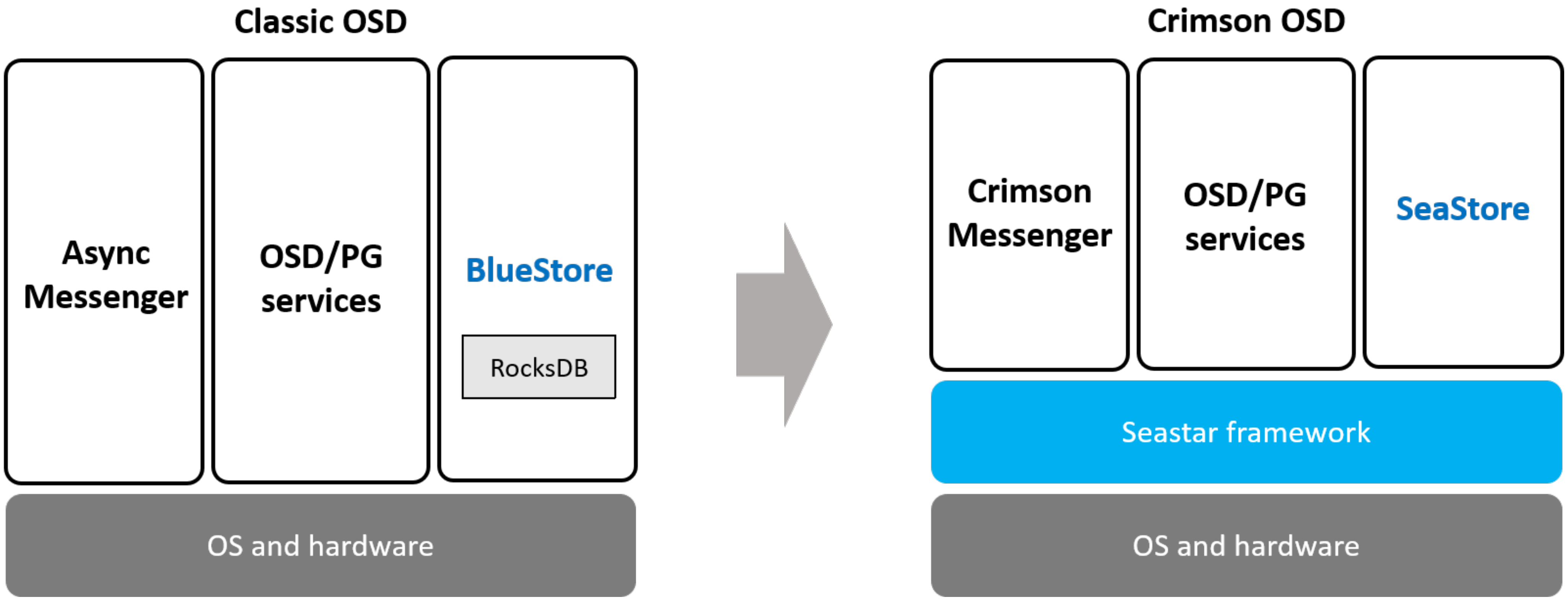
Crimson OSD Architecture



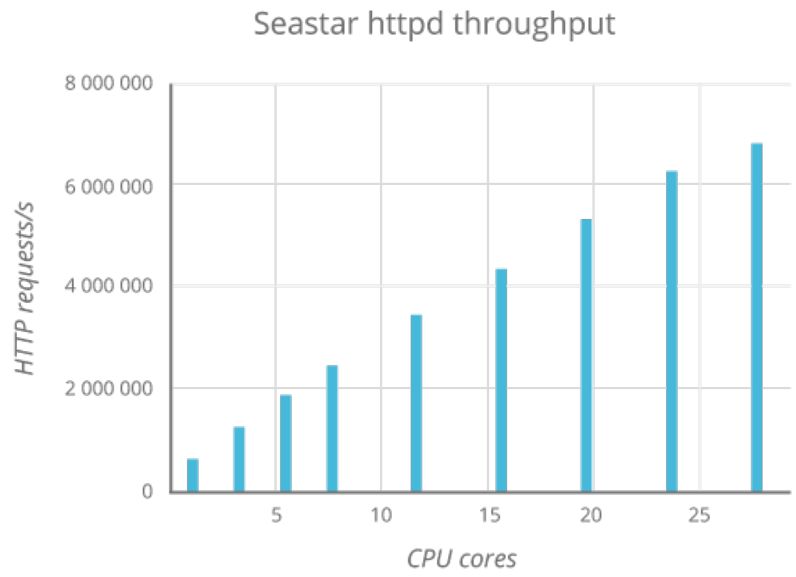
Shared-nothing (1 thread per core), cooperative userspace scheduling, run-to-completion, lockless, NUMA-aware memory allocator

~~Context switching, locking, shared resources~~

OSD Refactor with Seastar



SEASTAR.



- No need to reinvest the wheel
  - Seastar – a high-performance C++ framework
  - Shared-nothing design, 1 thread-per-core
  - Lockless NUMA-aware allocator, I/O abstraction, syscalls, event-center, …
  - Proven CPU scalability (Scylla, Pedis, Memcached)
  - Open source, Community backed
- Asynchronous Programming model
  - Explicit message passing across cores
  - Asynchronous: Callbacks Futures and Continuations

```
return sleep(5s).then([] {
  return read(4KB);
}).then([](buffer out) {
  return write(out);
}).then([]() {
  return flush();
});
```

Crimson Features

Core features functional (Ceph “Reef”)

- librados operations including snapshot support
- Log based recovery and Backfill
- RBD workloads on Replicated pools
- AlienStore (BlueStore), CyanStore (memory-based), SeaStore core
- Deployment via Cephadm

Ongoing (Ceph “S__”)

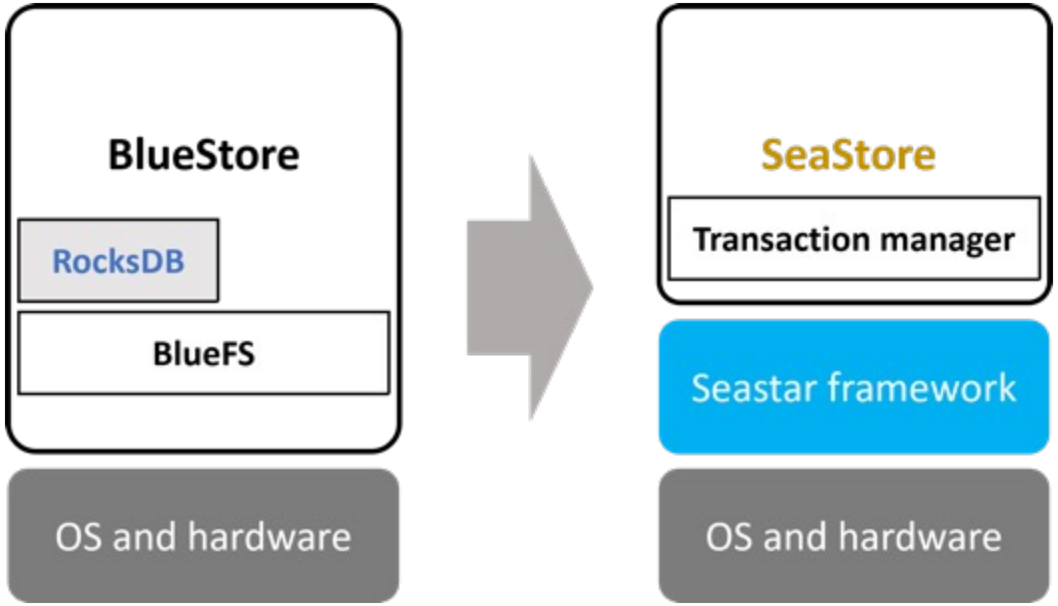
- Scrub, Erasure Coding
- Multi-shard support, stabilization, performance (RBD, RGW S3)
- SeaStore stabilization

Longer-term

- SeaStore heterogeneous storage support
- OSD PG Scaling, Ceph-FS

SeaStore

- New ObjectStore designed natively for Crimson threading/callback architecture
- Support emerging storage technologies (NVMe/ZNS)
- Designed to scale better than BlueStore



SeaStore Status - Current Features

Core features functional (Ceph “Reef”)

- Data read/write
- Indexing: object, address, metadata
- Transactions, background cleaning, …

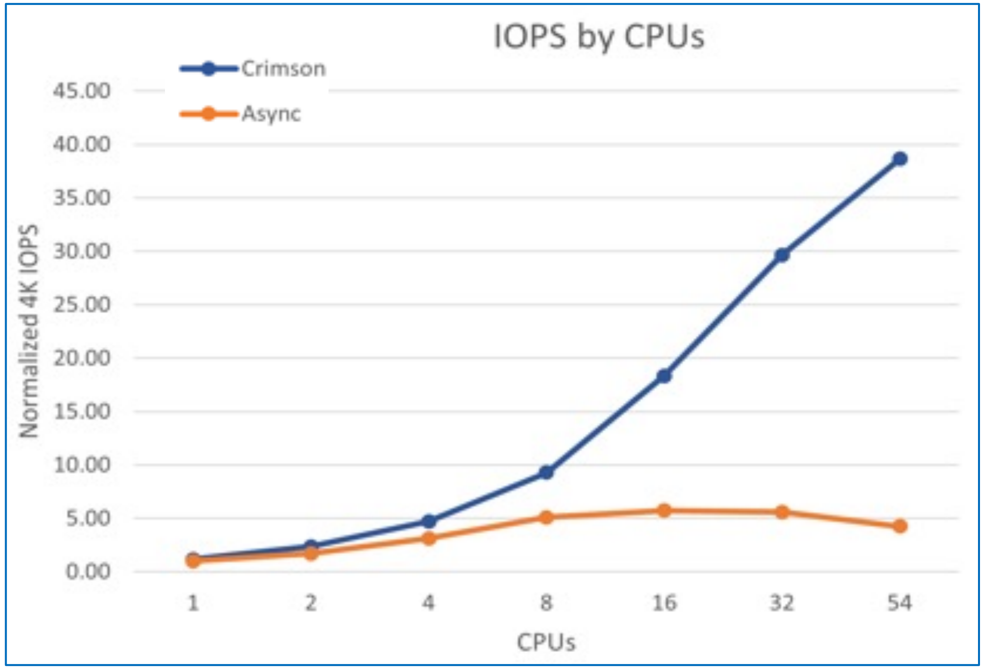
Ongoing (Ceph “S__” and beyond)

- Stability
- Device tiering
- Multiple-core run-to-completion
- Snapshot

Crimson Messenger Core Scalability (Reef)

Multi-shard Messenger

- Multi-shard Messenger ready, OSD WIP (target “S” release)
- Each thread working independently in the I/O path due to the shared-nothing design

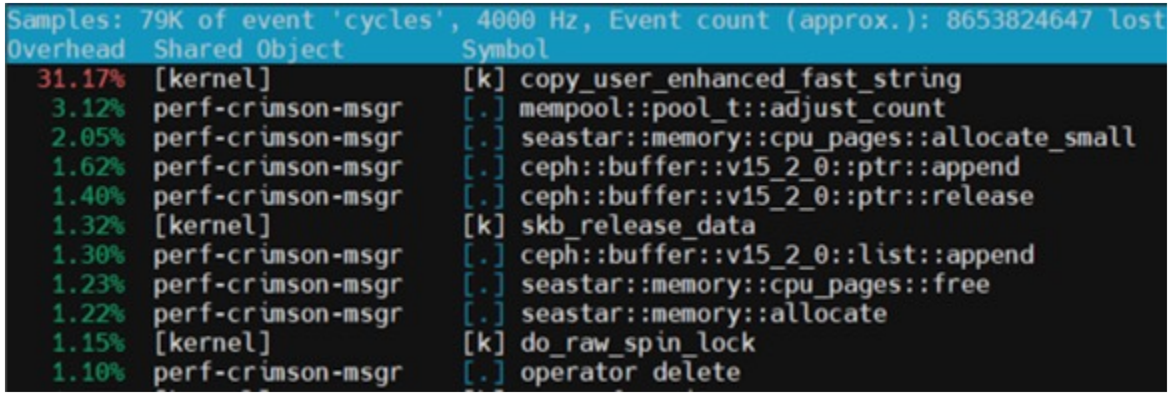


Crimson Messenger Core Scalability (Reef)

Multi-shard Messenger

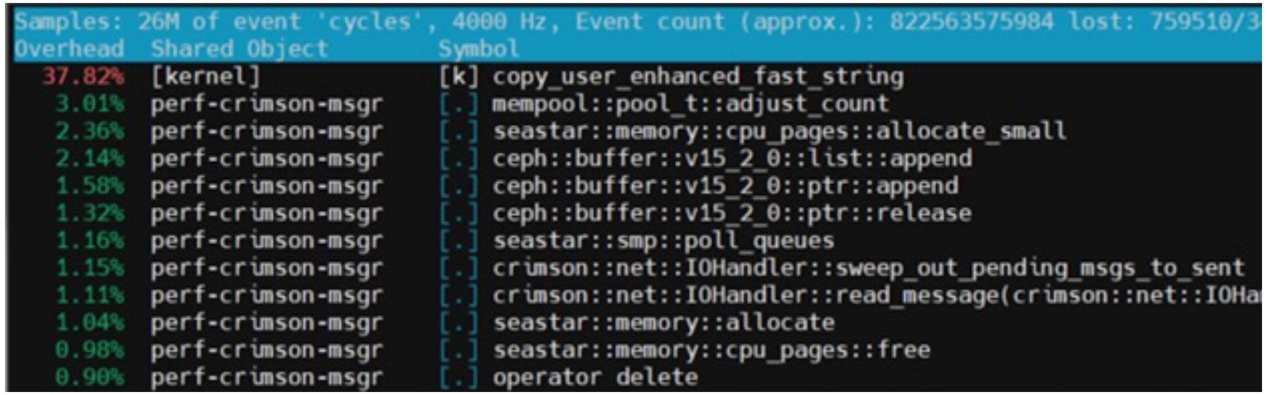
- Multi-shard Messenger ready, OSD WIP (target “S” release)
- Each thread working independently in the I/O path due to the shared-nothing design
- Good scaling trend (Performance hotspot pattern consistent from 1 to 100+ cores)

1 core



```
Samples: 79K of event 'cycles', 4000 Hz, Event count (approx.): 8653824647 lost
Overhead  Shared Object            Symbol
  31.17%  [kernel]                 [k] copy_user_enhanced_fast_string
   3.12%  perf-crimson-msgr        [.] mempool::pool_t::adjust_count
   2.05%  perf-crimson-msgr        [.] seastar::memory::cpu_pages::allocate_small
   1.62%  perf-crimson-msgr        [.] ceph::buffer::v15_2_0::ptr::append
   1.40%  perf-crimson-msgr        [.] ceph::buffer::v15_2_0::ptr::release
   1.32%  [kernel]                 [k] skb_release_data
   1.30%  perf-crimson-msgr        [.] ceph::buffer::v15_2_0::list::append
   1.23%  perf-crimson-msgr        [.] seastar::memory::cpu_pages::free
   1.22%  perf-crimson-msgr        [.] seastar::memory::allocate
   1.15%  [kernel]                 [k] do_raw_spin_lock
   1.10%  perf-crimson-msgr        [.] operator delete
```

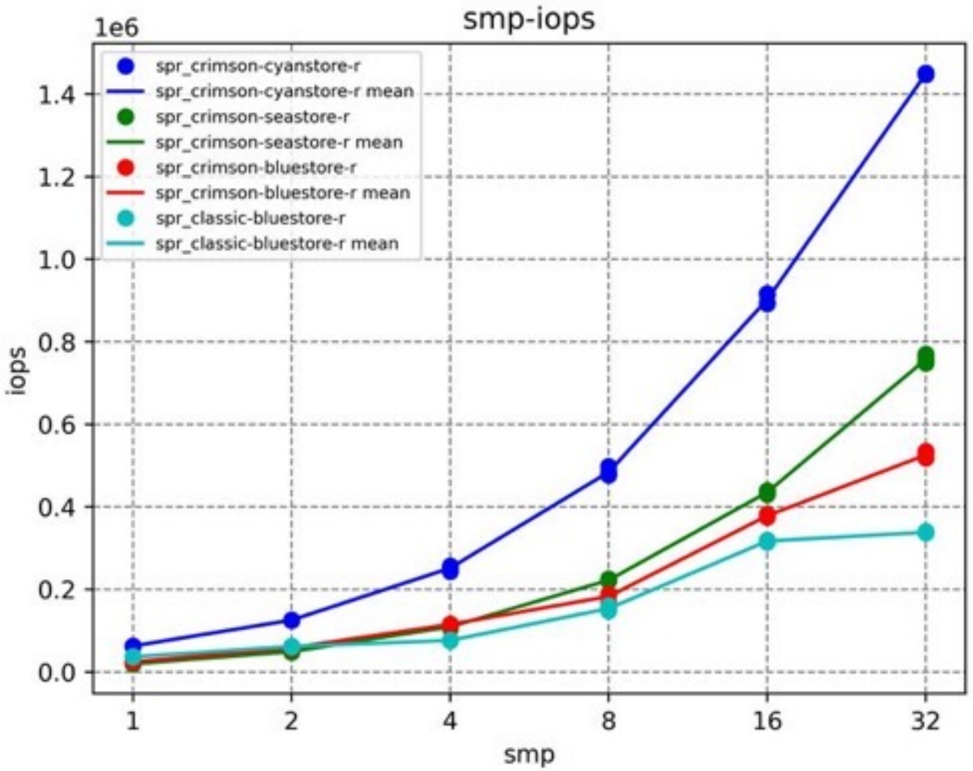
100+ cores



```
Samples: 26M of event 'cycles', 4000 Hz, Event count (approx.): 822563575984 lost: 759510/3
Overhead  Shared Object          Symbol
  37.82%  [kernel]               [k] copy_user_enhanced_fast_string
   3.01%  perf-crimson-msgr      [.] mempool::pool_t::adjust_count
   2.36%  perf-crimson-msgr      [.] seastar::memory::cpu_pages::allocate_small
   2.14%  perf-crimson-msgr      [.] ceph::buffer::v15_2_0::list::append
   1.58%  perf-crimson-msgr      [.] ceph::buffer::v15_2_0::ptr::append
   1.32%  perf-crimson-msgr      [.] ceph::buffer::v15_2_0::ptr::release
   1.16%  perf-crimson-msgr      [.] seastar::smp::poll_queues
   1.15%  perf-crimson-msgr      [.] crimson::net::IOHandler::sweep_out_pending_msgs_to_sent
   1.11%  perf-crimson-msgr      [.] crimson::net::IOHandler::read_message(crimson::net::IOHa
   1.04%  perf-crimson-msgr      [.] seastar::memory::allocate
   0.98%  perf-crimson-msgr      [.] seastar::memory::cpu_pages::free
   0.90%  perf-crimson-msgr      [.] operator delete
```

Crimson Performance – Core Scalability (Reef)

- Reads show good scaling in initial tests
  - Read IOPS scale linearly with cores
  - Good scaling with BlueStore/SeaStore
  - Architecture directionally correct
- Writes scale well with memory-backed Cyanstore
  - Crimson+ObjectStore write performance current optimization focus



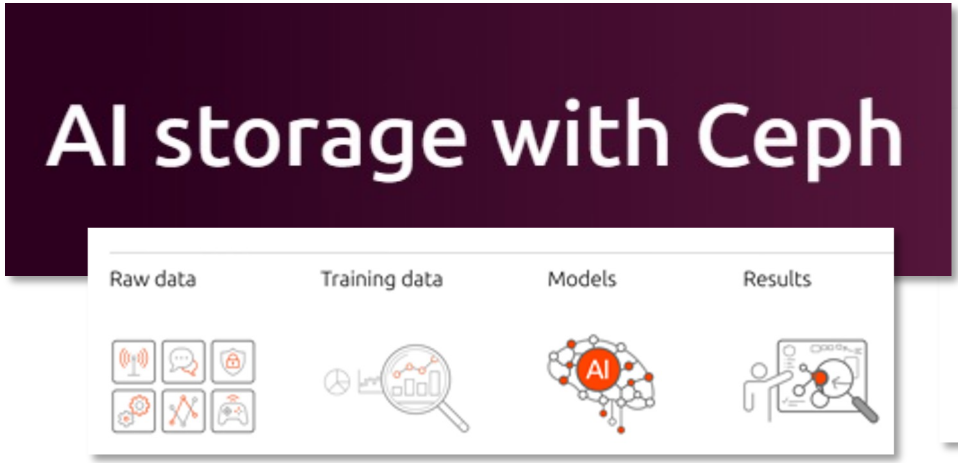
* Constrained scaling test with 1 OSD, 1 replica. Results may vary. `rados bench 4K randrd, depth=128, pgs=128, 30s; 1 OSD, 1 replica, 1~32 cpus`

Crimson Resources

Crimson Landing Page	https://ceph.com/en/news/crimson/
Project Documentation	https://docs.ceph.com/en/latest/dev/crimson/
Codebase	https://github.com/ceph/ceph/tree/master/src/crimson
Discussions	https://pad.ceph.com/p/crimson-weekly-meeting
Pull Requests	https://github.com/issues?q=is%3Apr+label%3Acrimson
SeaStore profiling	https://www.youtube.com/watch?v=SUJjZ9bjXJc

Storage of Choice for AI

Ceph: Leading Open-Source Scale-Out Solution for AI



AI storage with Ceph

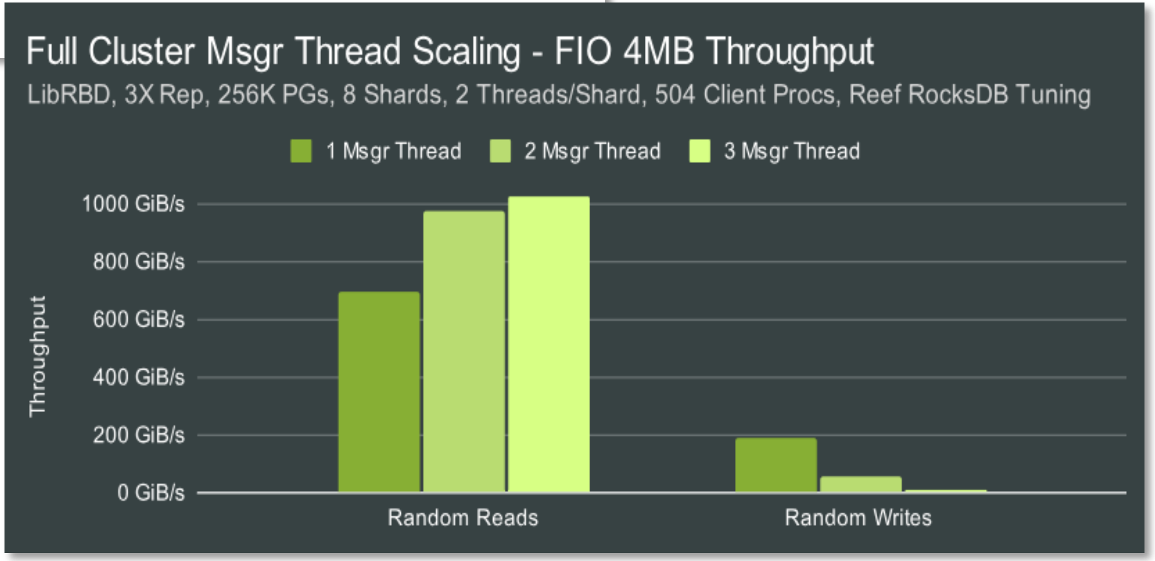
Nvidia DGX SuperPOD

Upto 1024 petaFLOPs (FP8)

>1 TBps storage throughput recommended

Ceph

A Journey to 1 TiB/s



Full Cluster Msgr Thread Scaling - FIO 4MB Throughput

LibRBD, 3X Rep, 256K PGs, 8 Shards, 2 Threads/Shard, 504 Client Procs, Reef RocksDB Tuning

- AI storage use cases
- Storage economics
- Performance considerations
- Storage security

Philip Williams, Mar 12, 2024
https://ubuntu.com/engage/ai-storage-with-ceph

Mark Nelson, Jan 19, 2024
https://ceph.io/en/news/blog/2024/ceph-a-journey-to-1tibps/

Ceph in the News…

Ceph for data lakehouses, generative AI

“Ceph is optimized for large single and multisite deployments and can efficiently scale to support hundreds of petabytes of data and tens of billions of objects, which is key for traditional and newer generative AI workloads.”

Gerald Sternagl, manager at IBM Storage Ceph

By Chris Mellor - January 30, 2024
[IBM touts Ceph for data lakehouses, generative AI – Blocks and Files]

Ceph as underlying AI data store

“In the Ceph world you roll in another 100 TB in a box, add it to the cluster, and off you go.” Ceph will automatically be able to use that.

“The WatsonX team are working closely with Ceph.” WatsonX being IBM’s generative AI platform.

Denis Kennelly, GM IBM Storage

By Chris Mellor - February 14, 2024
[IBM using Ceph as underlying AI data store – Blocks and Files]

Q&A

Thanks for Viewing this Webinar

- Please rate this presentation and provide us with feedback
- This webinar and a copy of the slides are available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webinar will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us @SNIACloud

Thank You!