# An Exploration into Object Storage for Exascale Supercomputers

Raghu Chandrasekar

# Agenda

- Introduction
- Trends and Challenges
- Design and Implementation of SAROJA
- Preliminary evaluations
- Summary and Conclusion

# Safe Harbor Statement

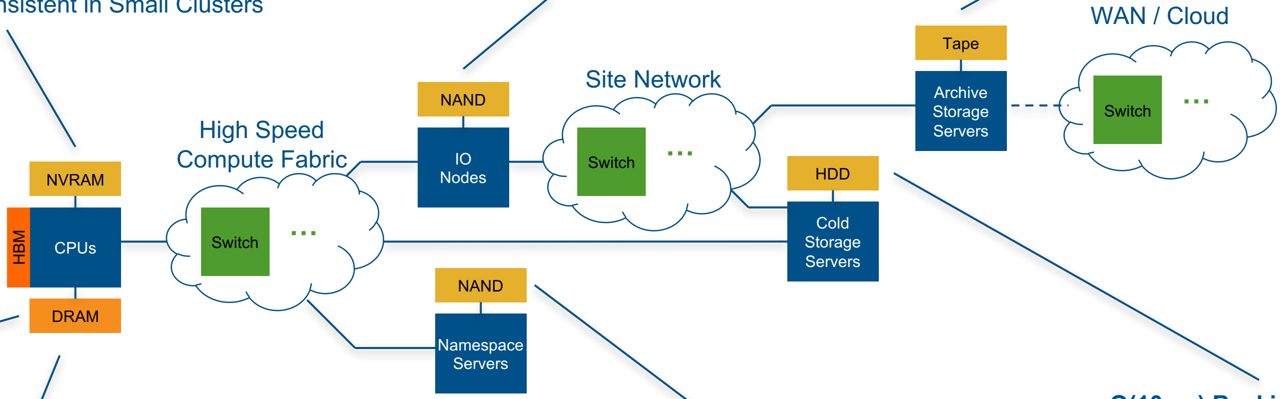
This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

# Storage Hierarchy Data Path Concepts

**O(1µs) Nonvolatile Storage**
*Private Scratch Namespace*
Relaxed POSIX or Key-Value API
Sharable Namespace (upon flush to backing store)
App-Controlled Caching; Close-to-Open Consistency
Inter-Node Cache-Consistent in Small Clusters

**O(100µs) Nonvolatile Storage**
*Sharable Across Computers*
Lightweight Object or POSIX Interface
Primary Resilient Random Storage
For HPC / Analytics / General

**Multi-Second Data Distribution & Protection**
*Multi-Site Reach*
Bulk Object Storage API
Cloud Bursting, Disaster Recovery, Archival



**O(5ns) In-Package Memory**
Customer Workload

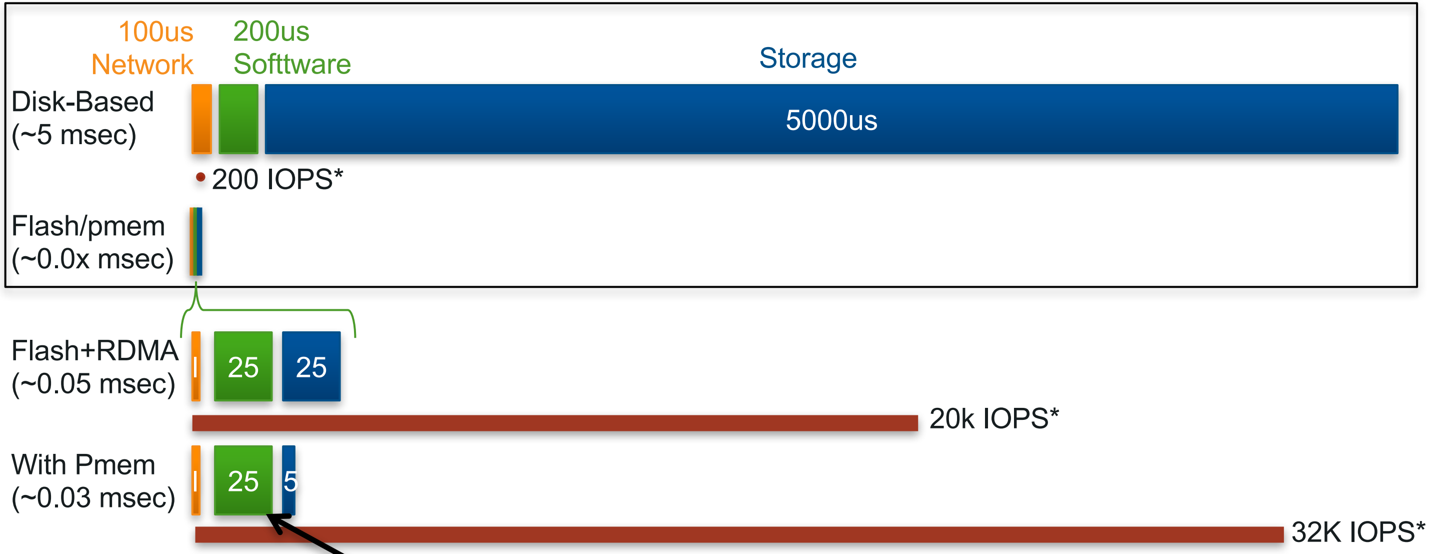
**O(50ns) DDR Memory**
Customer Workload

**Distributed Storage Client**
*Node-Local Cache & Working Memory*
Relaxed POSIX or Key-Value API
Local Cache Control
Sharding and Resiliency Controls

**Scalable Metadata Service**
*Serves Metadata to One Or More Tiers*
Provide Key-Value, POSIX,  namespace APIs
Attributes and Rich Metadata Structures
Map App Structures to One Another
Collections and Manifests of Large Data Sets

**O(10ms) Backing Store**
*Site-Wide Access*
Bulk Object/Cloud Storage API
High-9's Data Resiliency
Streaming Sequential Throughput

# Storage Media Latencies and IOPs

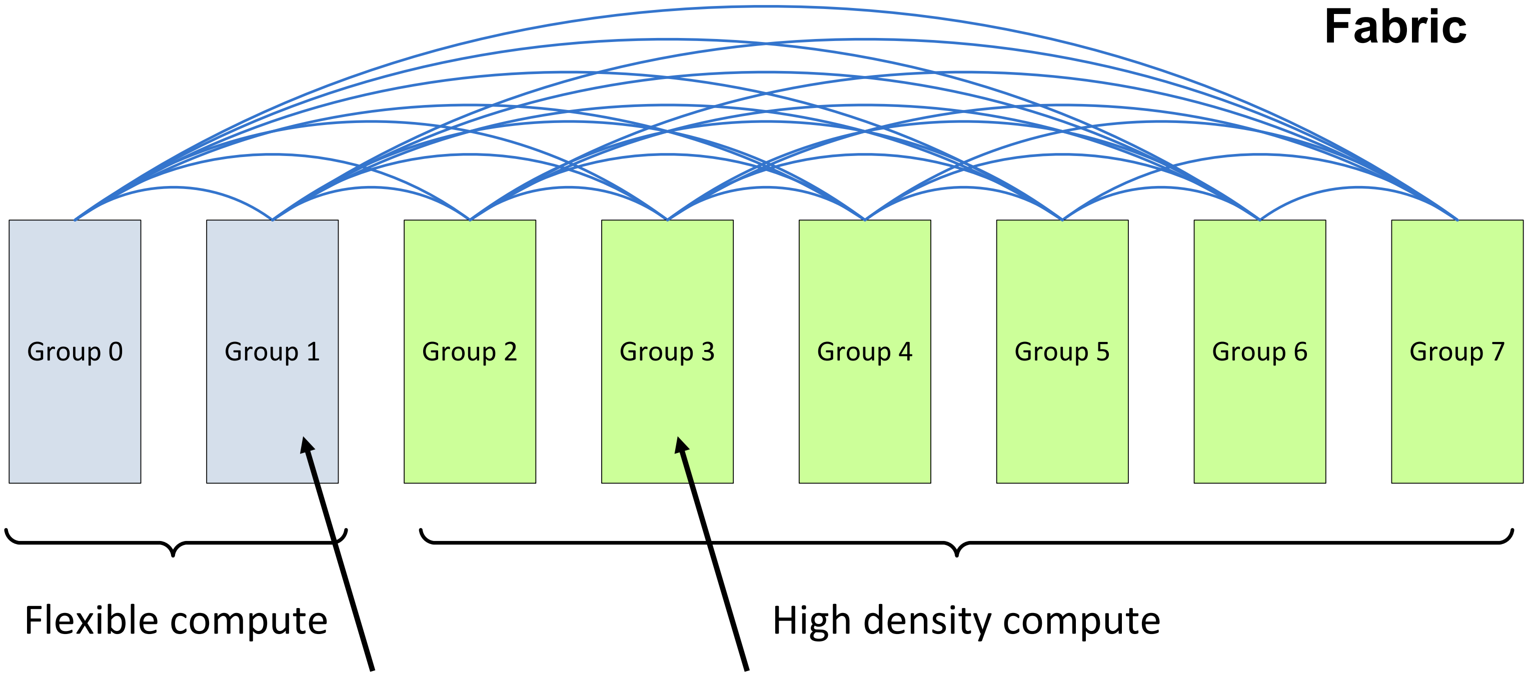


***Software becomes the largest fraction of latency when using persistent memory, even with 4x improved software efficiency***

* Max potential 1-thread random sector

# Cray Compute and Fabric Topology

High bandwidth Dragonfly Fabric



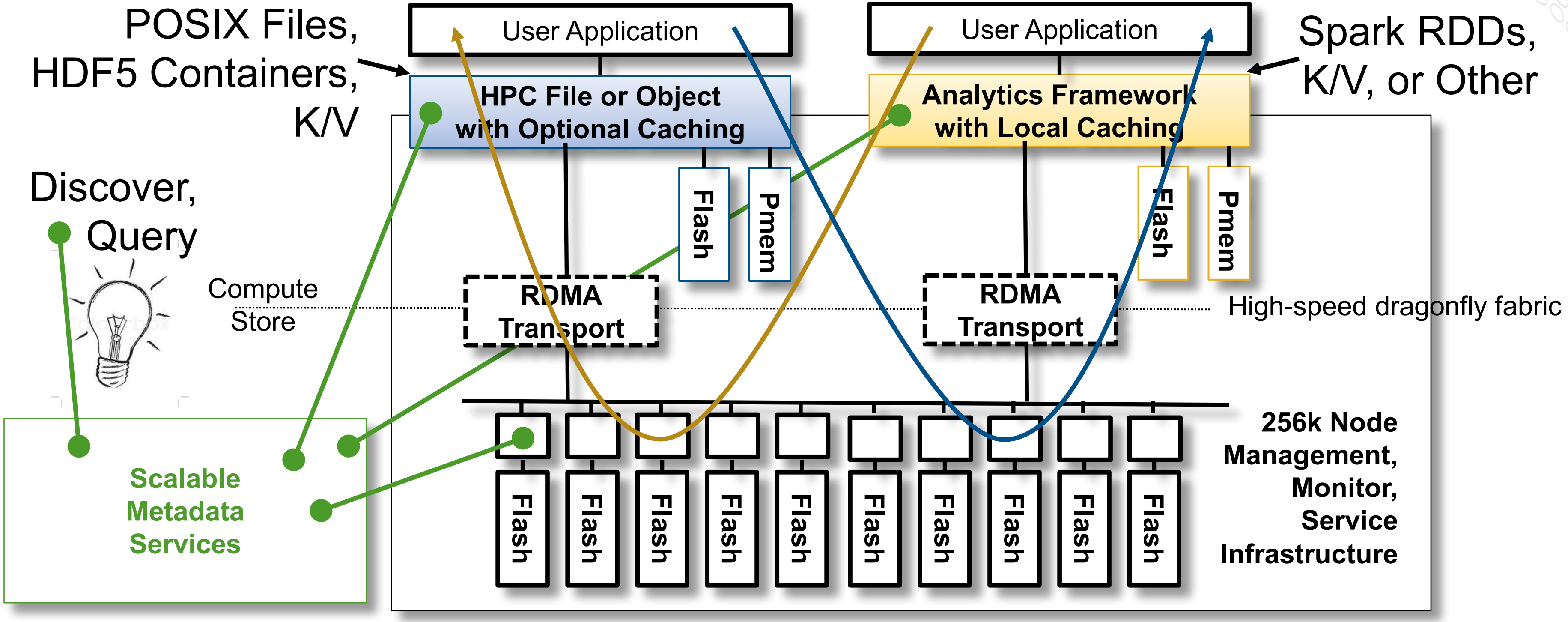
Flexible compute

High density compute

**Enclosure-Based Storage**
**Potentially 64k (or more) Devices**

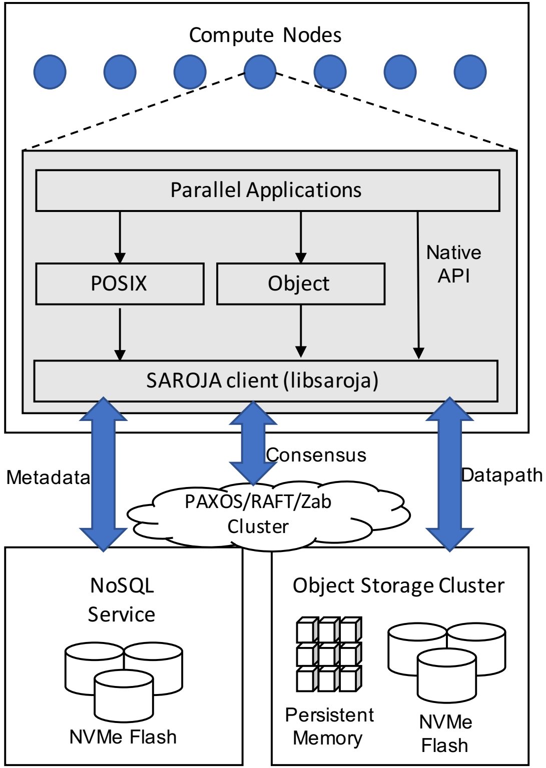
**Compute Node-Local Storage**
**Potential 256k Nodes**

# Analytics and HPC Software Convergence



# SAROJA Proof-of-Concept

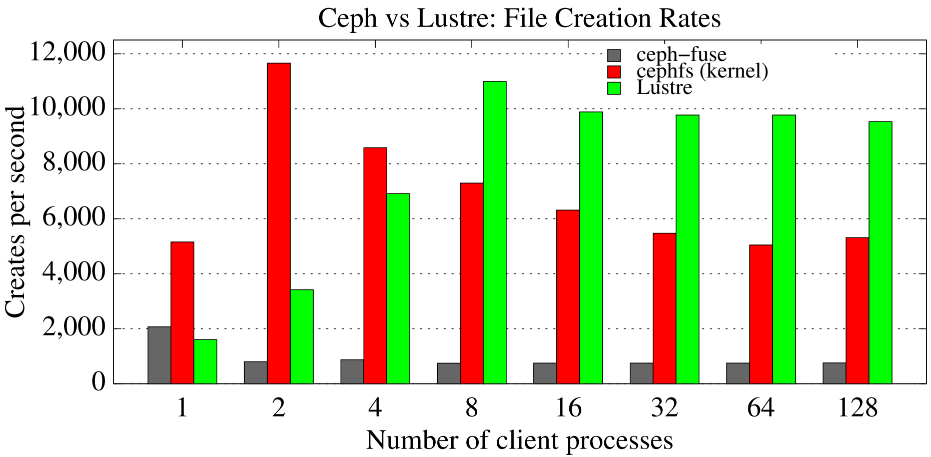
# Scalable And Resilient ObJect StorAge



# Preliminary Evaluations

# Metadata Evaluations

(Higher is better)

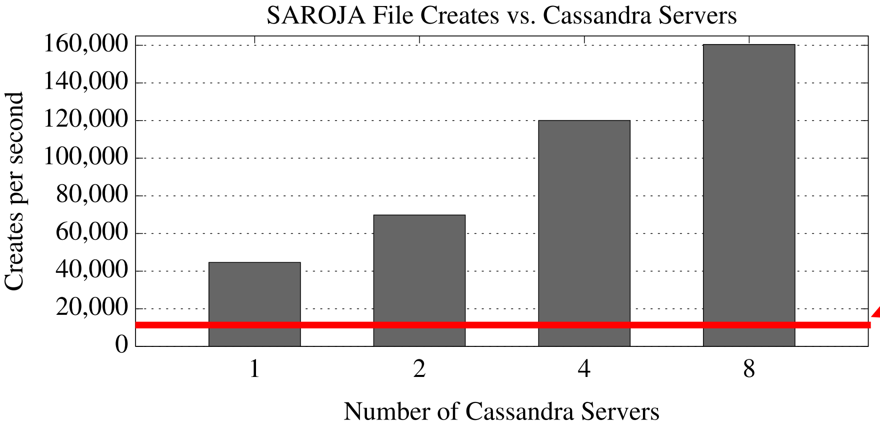


| | Ceph | Lustre |
|---|---|---|
| Software Version | v11.0.0 | v2.7.1 |
| Object Servers | 4 | 4 |
| Number of SSDs | 24 | 24 |
| Replication Factor | 1 | N/A |
| Number of MDS | 1 | 1 |
| Storage Backend | BlueStore | 1 OST-per-SSD |
| Fabric interface | IPoIB | IPoIB |
| Network driver | SimpleMessenger | sockets LND |

**Ceph POSIX support still has a long way to go**

# Metadata Evaluations

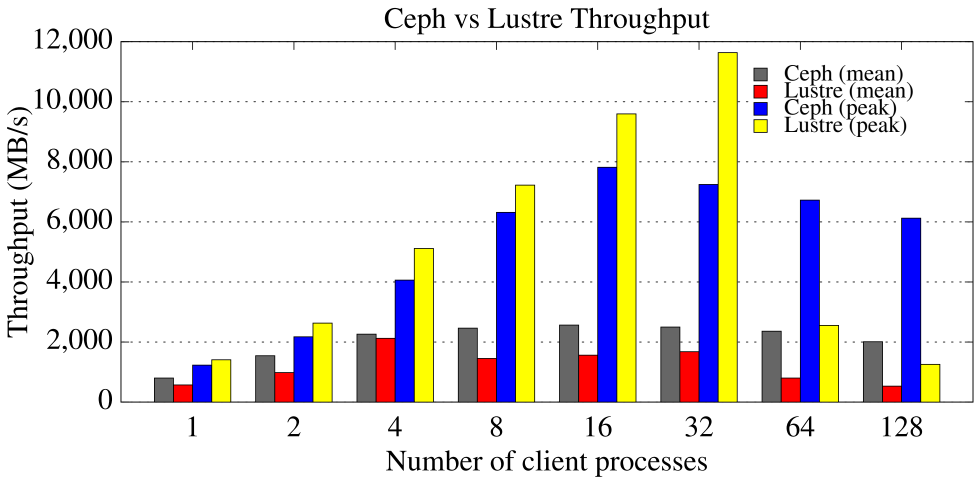
- POSIX over SAROJA
- 4480 MPI ranks
- 56 XC compute nodes
- 500 files/rank
- TCP over GNI
- Replication disabled



Peak Lustre file creation rate (w/o DNE)

**Scaling trends not ideal; but promising approach functionally**

# Data Path Evaluation



| | Ceph | Lustre |
|---|---|---|
| Software Version | v11.0.0 | v2.7.1 |
| Object Servers | 4 | 4 |
| Number of SSDs | 24 | 24 |
| Replication Factor | 1 | N/A |
| Number of MDS | 1 | 1 |
| Storage Backend | BlueStore | 1 OST-per-SSD |
| Fabric interface | IPoIB | IPoIB |
| Network driver | SimpleMessenger | sockets LND |

**Viable for use in the data path;**
**Plenty of opportunities for tuning**

# Summary

- Inflection point in storage system design
- Three-tier storage topology for supercomputers
- Promising early investigations with object storage tech
- Gradual transition
- Call for feedback

# Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, REVEAL, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

# Questions & Answers

Raghu Chandrasekar
raghu@cray.com