Some lessons learned from creating and using the Ceph open source storage system

Carlos Maltzahn, UC Santa Cruz
Presented at BCS Open Source Specialist Group, June 11, 2020
About me

Adjunct Professor, Step 4, Computer Science & Engineering, UC Santa Cruz

Founder & Director, Center for Research in Open Source Software (cross.ucsc.edu)

Co-Founder & Director, UCSC Systems Research Laboratory (SRL)

1999-2004: Performance Engineer, Netapp

Advising 5 Ph.D. students

Graduated 9 Ph.D. students, 9 M.S. students

people.ucsc.edu/carlosm

Current Research:

- Programmable Storage Systems (programmability.us)
- Big Data Storage & Processing
- Scalable Data Management
- Distributed Systems Performance Management
- Practical Falsifiable Research (falsifiable.us)

Past Research:

- Team processes in repositories [IJICIS’92]
- Network Intermediaries [SIGMETRICS’97]
- Automatic Behavioral Modeling of HDDs [MSST’14]
- Data Management Games [GamifIR’14]
Current Funding


NSF/CNS: Declarative Programmable Storage (with Peter Alvaro, 2018-2021)

NSF/CNS: Data Center Scale Programmable Storage (with Dirk Grunwald, CU Boulder, 2017-2020)


NSF/OAC: Big Weather Web: A Common and Sustainable Big Data Infrastructure in Support of Weather Prediction Research and Education in Universities (with 7 other Universities, NCAR, and one startup, 2015-2019)

CROSS: Skyhook: Programmable Storage for Databases (2016-)

CROSS: Eusocial Storage Devices (2019-)

CROSS: Black Swan: The Popper Reproducibility Platform (2019-)

CROSS: Mapping Datasets to Object Storage (2019-)

The picture can't be displayed.
**Ph.D. students, research faculty, and staff**

<table>
<thead>
<tr>
<th>5th Year</th>
<th>4th Year</th>
<th>3rd Year</th>
<th>1st Year</th>
<th>2nd Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jianshen Liu</td>
<td>Yiming Zhang</td>
<td>Aaron Chu</td>
<td>Saheed Adepoju</td>
<td>Esmaeil Mirvakili</td>
</tr>
</tbody>
</table>

**Research Staff**
- Dr. Kate Compton
- Dr. Ivo Jimenez
- Dr. Jeff LeFevre

**Research Faculty**
- Ivo Jimenez (UCSC)

**Graduated PhDs in previous years**
- Alexander Ames (2011, LLNL)
- Joe Buck (2014, Okera)
- Adam Crume (2015, Google)
- Latchesar Ionkov (2018, LANL)
- Michael Sevilla (2018, TidalScale)
- Andrew Shewmaker (2016, OpenEye)
- Dimitrios Skourtis (2014, IBM Almaden)
- Noah Watkins (2018, vectorized.io)
Sage Weil & the beginnings of Ceph

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Sage’s summer internship at LLNL, SC’04 paper</td>
</tr>
<tr>
<td>2005</td>
<td>“Ceph” summer project -&gt; thesis topic</td>
</tr>
<tr>
<td>2006</td>
<td>OSDI’06 and SC’06 papers</td>
</tr>
<tr>
<td></td>
<td>● Published source code licensed under LGPL-2.1</td>
</tr>
<tr>
<td>2007</td>
<td>Sage’s Ph.D. defense</td>
</tr>
<tr>
<td>2008</td>
<td>Sage continues to work on Ceph</td>
</tr>
<tr>
<td>2010</td>
<td>Ceph Linux kernel client in 2.6.34</td>
</tr>
<tr>
<td>2011</td>
<td>Inktank Storage startup</td>
</tr>
<tr>
<td>2014</td>
<td>Red Hat acquires Inktank for $175m</td>
</tr>
</tbody>
</table>
Users & Contributors in Industry

A wide range of companies are making long-term strategic bets on Ceph, and that number is growing!

Here are a few users to give you an idea of the breadth of use that Ceph is seeing in the ecosystem today.

Source: ceph.io
A Transformative Gift by Alumnus Sage Weil

- $500K Gift + $500K UC Matching for The Sage Weil Presidential Chair in Open Source Software
- $2M gift to establish CROSS and support open source software research
Sage Weil & the beginnings of Ceph

- 2004: Sage’s summer internship at LLNL, SC’04 paper
- 2005: “Ceph” summer project -> thesis topic
- 2006: OSDI’06 and SC’06 papers
  - Published source code licensed under LGPL
- 2007: Sage’s Ph.D. defense
- 2008: Sage continues to work on Ceph
- 2010: Ceph Linux kernel client in 2.6.34
- 2011: Inktank Storage startup
- 2014: Red Hat acquires Inktank for $175m
• Bridges gap between student research & open source projects
• Funded by Sage Weil endowment & corporate memberships

Teach students on how to productively engage in open source communities
Fund high-impact research with plausible path to successful open source projects
Incubate developer communities around research prototypes
The picture can't be displayed.
Advisory Committee

**Doug Cutting** (Chief Architect, Cloudera)
Founder of numerous groundbreaking open-source projects, including Lucene, Nutch, Avro, Hadoop

**Karen Sandler** (Executive Director, Software Freedom Conservancy)
Former executive director of GNOME Foundation, former general counsel of Software Freedom Law Center

**Nissa Strottman** (VP, Technology, IP, Innovation, Strategic Partnerships, VISA),
Years of experience in legal and M&A in the open-source software space.

**Sage Weil** (Chief Architect of Ceph, Red Hat)
Creator of WebRing, founder of DreamHost, founder and CTO of Inktank, recipient of O’Reilly Open Source Award

**James Davis** (Professor of Computer Science and Engineering)
He was Founding Director of UCSC Center for Entrepreneurship, sits on advisory councils for startups and nonprofits. Interested in research applying technology to address global social issues, and has developed an award winning course around this theme.
Why these companies?
Leverage Ceph to create and shape new markets!

Industry Advisory Board

$100,000 / year / member
Operational Model

- Modelled after NSF’s I/UCRCs
- Adds open-source software focus
- Sustained through membership fees
- **Call for research and incubator proposals twice a year**
  - Topics of interest informed by industry members
- **Review of projects twice a year**
  - Project selection for funding informed by industry members

<table>
<thead>
<tr>
<th>Spring</th>
<th>Fall</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early January</td>
<td>May</td>
<td>Topics of interest input</td>
</tr>
<tr>
<td>Mid January</td>
<td>June</td>
<td>Call for proposals</td>
</tr>
<tr>
<td>Mid February</td>
<td>Begin September</td>
<td>Proposals due</td>
</tr>
<tr>
<td>End February</td>
<td>Mid September</td>
<td>Greenlighting</td>
</tr>
<tr>
<td>Mid March</td>
<td>Begin October</td>
<td>IAB meeting</td>
</tr>
</tbody>
</table>

**Diagram: CROSS**

- **Research**
- **Incubator**
- **Open-Source Software Education & Practice**
- **Open-Source Software Experts**
- **University Research**
  - Undergrad
  - Graduate
  - Postdoc

**Industry Advisory Board Review**

- Topics of interest input: Early January, May
- Call for proposals: Mid January, June
- Proposals due: Mid February, Begin September
- Greenlighting: End February, Mid September
- IAB meeting: Mid March, Begin October
Postdocs building dev communities for their research prototypes

Seeded by students working on term, senior, master, or PhD projects

Graduated with Ph.D. and is well-published expert

Starts out with a significant code base from Ph.D. project

Leverages at least one well-established OSS community

Wants to become an OSS leader

Community seeding via “Research Experience” Programs:

- Tap into pool of students who need project topics
- CROSS as Google Summer of Code Mentor Organization
- Great community management training
- Great driver for community infrastructure

Evaluation metric: number of contributors from number of organizations

Exit: when external funding becomes available or project fails reviews

Expected runtime: 2-4 years

Kate Compton: Tracery 2 & Chancery – Getting poets to program AI
tracery.io

Ivo Jimenez: Black Swan – The Practical Reproducibility Platform
falsifiable.us

Jeff LeFevre: SkyhookDM – Programmable Storage for Databases
skyhookdm.com
Incubator creating a market for open source project ideas
Matching mentors, students, and funders
Cutting-edge research projects with plausible paths to successful open source software projects

**Research Symposium**

Showing off student work at CROSS and other UC Santa Cruz research programs

Annual 2-day event with 2 tracks of program and “Systems Oktoberfest”, next event: Oct 6-9, 2020*

Centers technical program around current CROSS research and incubator projects

Shows off student work and research programs

Establishes interested communities of students, industry, government, and faculty

Located at Baskin School of Engineering on UC Santa Cruz campus

cross.ucsc.edu/symposium

---

**Requirements**

Addresses a fundamental research question

*Is advised by UC Santa Cruz faculty

*Is not required to create any software

Opens a plausible path to open source software that might be widely adopted

*Has completed coursework required by UC Santa Cruz Ph.D. program

---

**Graduated**

Ivo Jimenez (now incubator fellow): Popper - Practical Falsifiable Research (Advisor: Carlos Maltzahn)

Noah Watkins (vectorized.io): Zlog - Distributed Shared-log for Software-Defined Storage (Advisor: Carlos Maltzahn) *

Michael Sevilla (TidalScale): Mantle - A Programmable Metadata Load Balancer for the Ceph File System (Advisor: Carlos Maltzahn) *

Brendan Short: Strong Consistency in Dynamic Wireless Networks for Better Navigation of Autonomous Vehicles (Advisor: Ricardo Sanfelice)

---

**Research Fellows**

Xiaowei Chu: Mapping Datasets to Object Storage (Advisor: Carlos Maltzahn)

Akhil Dixit: CAvSAT - A System for Query Answering over Inconsistent Databases (Advisor: Phokion Kolaitis)

Aaron Hunter: OSAV – Autonomous Vehicle Controller (Advisor: Gabriel Elkaim)

Jianshen Liu: Eusocial Storage Devices (Advisor: Carlos Maltzahn)

Sheng Hong Wang: Live Hardware Development (LiveHD) (Advisor: Jose Renau)

---

*New!*

Kevin White: Strong Consistency in Dynamic Wireless Networks for Better Navigation of Autonomous Vehicles (Advisor: Ricardo Sanfelice)
What really works in CROSS

- **“Ceph as catalyst”:** Engagement of industry in research prototypes via open source
  - Curate and market open source portfolio to attract talent and industry funding
- **Incubator projects:** Engagement of top postdoc talent in incubating developer communities
  - Maintain advisory committee of world-class open source software experts
- **Research Experience:** Engagement of students to seed developer communities in incubator
  - Promote education to productively engage in open source communities
  - Help produce students with cutting-edge skills that are immediately marketable
- **Industry Practitioner-in-Residence:** Engagement in student mentorship and industry membership drive
  - Engage with government, industry, academia, and foundations via “standardization incubator”
- **Call for proposals:** Topics of interest by industry, research proposals by students/faculty, incubator proposals by postdocs
  - Match focused groups of funders, faculty, mentors, and students
Ceph as catalyst

• As adoption of Ceph goes up, ...
  • Value of Ceph-based research increases
    • Example of CMU’s recent SOSP’19 paper on Ceph’s BlueStore
  • Job market value of students increases
  • Funding opportunities multiply
    • Example of NSF IRIS-HEP, NSF funding for programmable storage

• Ceph as open source software provides framework for new standards
  • CROSS as a “standards incubator”
  • Component industry is too low-margin to afford research labs, must rely on academia
Universities need to value open source

• Linux, Drupal, Spark, LLVM, Ray, Ceph, BerkeleyDB, ...
• Long list of valuable projects that came out of universities
• Actual impact on society – not counted.
• Opportunities for mentorship, funding, industry relations – mostly ignored.
  • Research groups are using open source software for teaching and prototyping
• Software itself not considered academic contribution
  • Technology transfer offices often don’t know how to value of open source
• Time to create Open Source Program Offices for universities!
  • Contact me if you are interested!
The picture can't be displayed.
Ceph Overview

RADOS: Storage Servers

Clients
Ceph: Storage servers act collectively
Ceph: new storage service *by server*

**Name:** XYZ  
**API:** Read, Write

**Scope of method call:** local object

RADOS Storage Service

Clients
Ceph: new storage service by server

Name: XYZ
API:
Read
Write
Foo
Bar

Scope of method call: local object

Custom methods using Ceph CLS (aka custom object classes)

Clients

RADOS Storage Service
Ceph: new storage service by server

RADOS Storage Server

RADOS Storage Service
Ceph: new storage service by client

Scope of service: sets of objects
Ceph: new storage service by client

Scope of service: sets of objects
Ceph: new storage service by client

Scope of service: sets of objects
CephFS: File Mapping

Strategies:

**Striping**: e.g. stripe unit, stripe count, object size

**Object naming**: e.g. prefix = filename, suffix = seq#

**Object placement**: e.g. CRUSH(hash(objectname))

MDS provides striping and object naming
RADOS provides hash, CRUSH, and clustermap

- Primary copy replication or erasure coding
- Writes: atomic+transactional
- Given object name: Client can calculate location
- Partial reads/writes of objects
- Automatically redistributes data during failures/add/remove servers
SkyhookDM: Tabular Data

Table (schema, operations)

Object naming • Striping (here: n, 1, n)

Mapping data and operations

hash

CRUSH

Strategies:

**Striping:** e.g. stripe unit, stripe count, object size

**Object naming:** e.g. prefix = filename, suffix = seq#

**Object placement:** e.g. CRUSH(hash(objectname))

MDS provides striping and object naming
RADOS provides hash, CRUSH, and clustermap

- Primary copy replication or erasure coding
- Writes: atomic+transactional
- Given object name: Client can calculate location
- Partial reads/writes of objects
- Automatically redistributes data during failures/add/remove servers
What is SkyhookDM?

An object “class” for Ceph
- No upstream modifications required
- Inherits Ceph’s properties now and in the future
- Can use all other object extensions
- **Not a database**

Storing *tabular* data in objects
- Columnar data:
- Row data: FlatBuffers

Object read/write operations
- Select, Project, Aggregate
- Create, append rows/columns
- Indexing
- Intra- & inter-object transformations
Thanks to:

Thanks to our sponsors:
Kioxia, Micron, SK Hynix, Seagate, Western Digital, Huawei, Samsung, Fujitsu

Contact:
Carlos Maltzahn, carlosm@ucsc.edu
cross.ucsc.edu