Carlos Maltzahn

carlosm@ucsc.edu

11/22/16

Baskin Engineering UC SANTA CRUZ

World-class research in

- Big Data Storage & Processing
- Scalable Data Management
- Distributed Systems Performance Management

Baskin Engineering UC SANTA CRUZ

Carlos' Background

- **Adjunct Professor**, Computer Science Department, UC Santa Cruz
- **Director**, UCSC Systems Research Laboratory (SRL)
- **Director**, UCSC/LANL Institute for Scalable Scientific Data Management (ISSDM)
- **Director**, Center for Research in Open Source Software (CROSS)

- 1999-2004: **Performance Engineer**, Netapp

- **Advising** 6 Ph.D. students.
- **Graduated** 5 Ph.D. students
- **I do this 100% of my time!**

Baskin Engineering UC SANTA CRUZ

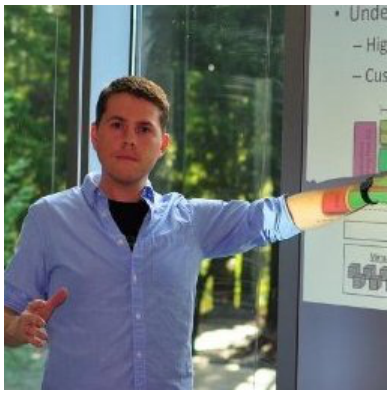
- Current Research
  - High-performance ultra-scale storage and data management
  - End-to-end Performance management and QoS
  - Network Intermediaries
- Other Research
  - Data Management Games
  - Information Retrieval
  - Cooperation Dynamics

Current Ph.D. students and staff at the SRL (E2.375)

Ivo Jimenez

Dr. Jeff LeFevre

Michael Sevilla



Noah Watkins

Jianshen Liu

Reza NasiriGerdeh

Yiming Zhang

Baskin Engineering UC SANTA CRUZ

Some SRL Success Stories

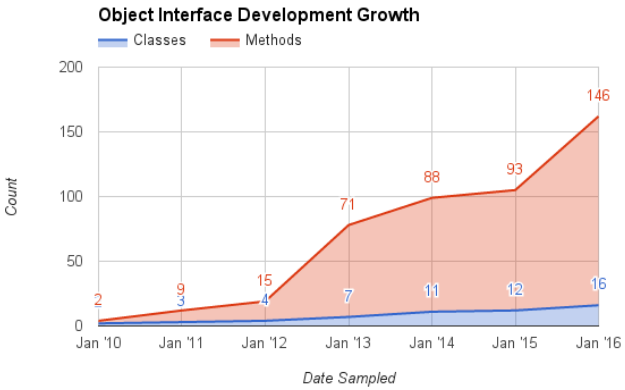
- **Ceph**
  - Consolidates storage tiers for Object, Block, and File
  - Yahoo! uses Ceph for Flickr and Cloud Object Store.
- **Fahrrad**
  - Robust efficient storage QoS (in NetApp's Data OnTAP)
- **RBED (Rate-Based Earliest Deadline)**
  - Guaranteed real-time scheduling (in Microsoft/ETH Barrelfish)
- **High-quality research and publications**
  - Recent Best Papers at SIGMOD, HPDC, ECRTS, and RTSS
- **Alumni: 4 Postdocs, 13 PhDs, 22 MSs**
  - Placements at IBM Research (×2), Symantec Research Lab (×2), MIT Lincoln Lab, NetApp, Yahoo!, Inktank, Disney, Lawrence Livermore National Lab, Los Alamos National Lab, TidalScale, …

SRL Research Areas (I)

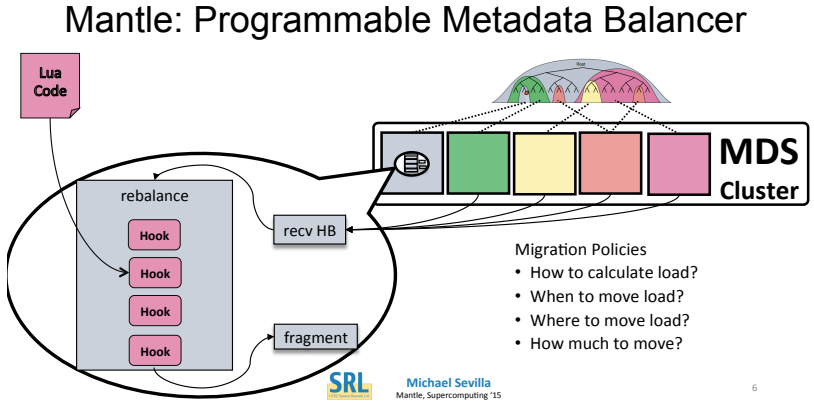
programmability.us

Programmability in Storage Systems

- Storage object interfaces
- Load balancing of metadata service



noahdesu.github.io



Sevilla SC'15

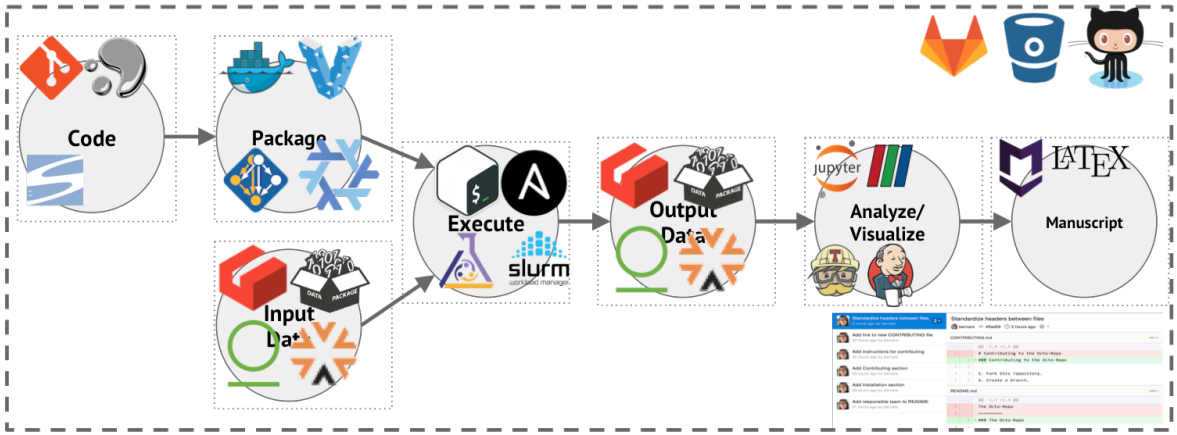
Baskin Engineering UC SANTA CRUZ

SRL Research Areas (II)

falsifiable.us

Reproducibility in Systems

Popper
falsifiable.us



Baskin Engineering UC SANTA CRUZ

Jimenez VarSys'16

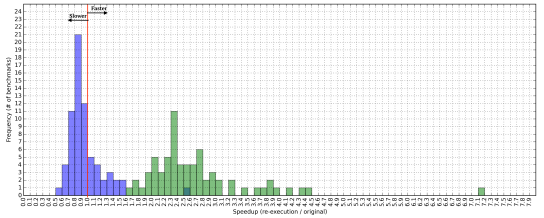


Fig. 2. [source] Histograms for two variability profiles. Each data point in a histogram corresponds to the performance speedup/slowdown of a `stress-ng` CPU method that a machine has with respect to another one. For example, in the T/B histogram (green), the speedup caused by the architectural improvements of machine T causes 11 stressors to have a speedup within the `(2.3, 2.4]` range over machine B.

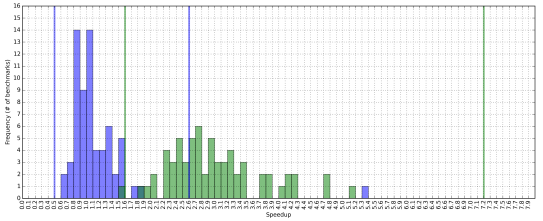


Fig. 3. [source] Histogram for T'/B and T/B profiles. The data points come from the following benchmarks: `STREAM`, `cloverleaf-serial`, `comd-serial`, `sequoia` (`amgmk`, `crystalmk`, `irsmk`), `c-ray`, `crafty`, `unixbench`, `stress-ng` (string, matrix, memory and cpu-cache). Vertical lines denote the limits of the predicted variability range (Figure 2), obtained from executing `stress-ng` CPU stressors. Points outside the predicted line correspond to `STREAM`. The rightmost point for the unconstrained (green) histogram is not shown to improve the readability of the figure; it lies on the 14x bin.

SRL Research Areas (III)

Distributed Systems Performance Management

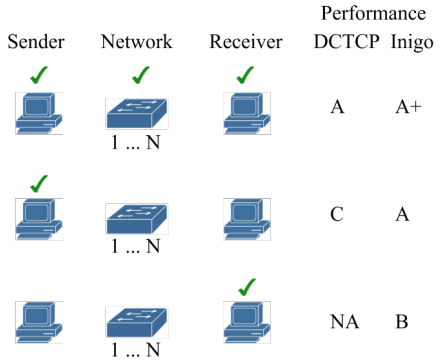


Figure 1: **Inigo's latencies are up to** 1.3× **better than DCTCP, the best deployable solution, when all components of a network are properly configured (green check). Inigo's sender-only mode is up** 42× **better than DCTCP's corresponding failure mode, according to fairness, bandwidth, and latency indices; and Inigo can also offer improvements when only the receiver is configured.** Letter grades are relative to a C for Reno-level performance.

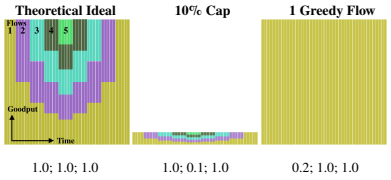
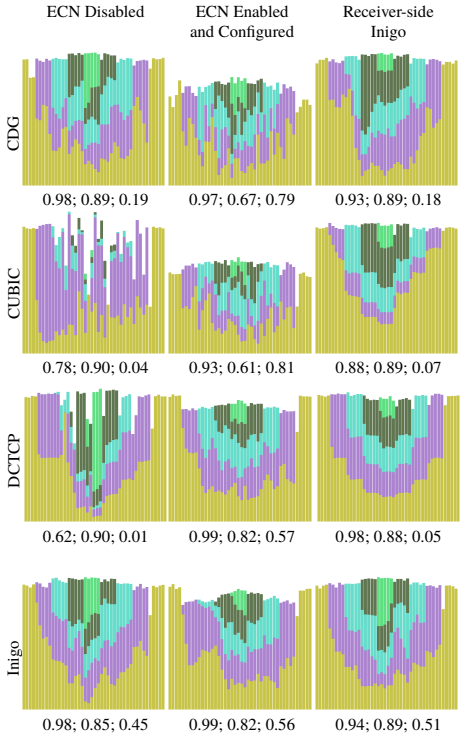


Figure 3: **Theoretical Ideal, 10% Capped, and 1 Greedy Flow examples.** Stacked Goodput vs. Time, five converging flows. Indices below: (1) Jain's Fairness Index; (2) Goodput Index; and (3) Latency Index of the Smoothed Round-trip-time distribution. Higher is better, and 1.0 is ideal. See beginning of Section § 5 for explanations.



Shewmaker ICCCN'16

SRL Research Areas (IV)

Scalable Data Management in Genomics

- 1,000,000 genomes
- Zetabytes
- International sharing
- Key to cancer, …

Scalability
- Size
- Geographic distance
- Administrative domains

Smart storage devices

?

Smart phones

Baskin Engineering UC SANTA CRUZ

SRL Summary

- Programmability of Storage Systems
- Reproducibility in Systems
- Distributed Systems Performance Management
- Scalable Data Management in Genomics

We want you to be successful!

- Finish your class work quickly and well
  - Use class projects as first baby steps into trying out new things!
- Finding a research problem:
  - "What's the fundamental challenge?"
  - "Can I falsify my assumptions?"
  - "Do I really like to think about that?"
    - Develop good taste for problems!
- Have a client for the problem:
  Problem solver ≠ Problem expert
- Start with dumbest approach and surprise yourself
  - Assumptions = Blind spots
- Look for solutions in multiple disciplines
  - "Foreign" concepts = Diversity of perspectives
- Build functioning prototypes:
  Multiple subsystems ≠ System

Enjoy writing and talking about technologies that few or none have explained before!

"*The distance from who we are to who we would like to be is often shorter than we think. But almost always farther than the couch.*"
[Henri, The Chat Noir, 3/18/2013]

Baskin Engineering UC SANTA CRUZ

Use your summers!

- Summer internships
  - Companies have started recruiting for next summer!
  - Consider a national lab for at least one summer: LANL, Sandia, LLNL, LBNL
    - Many if not most fundamental technological innovations happen there!
  - Consider source of funding beyond summer
- Participate in open source software projects
  - Professional communities of programmers
  - Start working on your “github portfolio”

cross.ucsc.edu

Baskin Engineering UC SANTA CRUZ

The Role of Open Source Software in Research & Education

Education

- The (only) examples of industry-strength code
- Open source communities include professional and experienced programmers
- Github projects become an important part of resume
- Coding is part of job interview: Google, Microsoft, Facebook, LinkedIn, Twitter, Amazon, Zynga, Dropbox, …
- Why does someone graduated with a Ph.D. in CS has to prove coding skills?

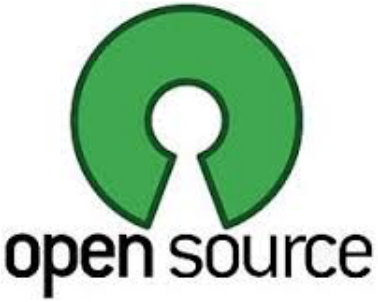
Research

- Essential for computer science
  - Enabler of scientific results
  - Enabler of reproducibility of scientific results
- We produce students and papers
- What about all those software prototypes?

Baskin Engineering UC SANTA CRUZ

The Opportunity of Open Source Software in Research & Education

Graduation!



Before

After

Baskin Engineering UC SANTA CRUZ

The Opportunity of Open Source Software in Research & Education

Before

Graduation!

After

Baskin Engineering UC SANTA CRUZ

Successful OSS projects

- Attract new talent
  - Help UCSC in recruiting very talented students
  - University student career as a path to OSS leadership
- Used as research & education platform
  - Leverage past systems research and make results reproducible
  - Useful and usable as a tool for systems education
- Create OSS leaders who know how to get systems built
  - Know where to find the right tools & technologies
  - Much more leveraged in their value to industry!

Baskin Engineering UC SANTA CRUZ

The Opportunity of Open Source Software in Research & Education

Graduation!

open source

Before

After

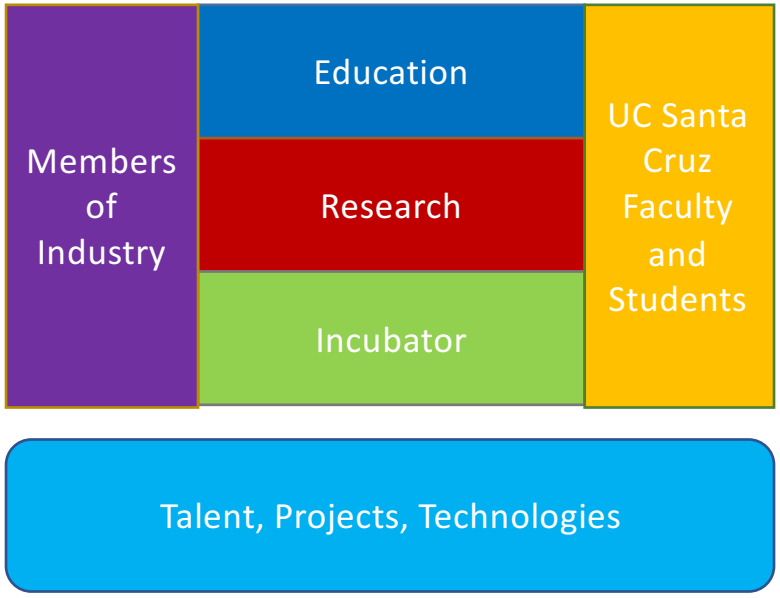
?

Baskin Engineering UC SANTA CRUZ

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

- **Bridges gap between student research & open source projects**
- Funded by Sage Weil endowment & corporate memberships
- Attracts talented students and employees who acquire skills in open source software

- Educate the next generation of OSS leadership
- Leverage OSS culture in UCSC research
- Incubate work beyond graduation to reach critical mass



Baskin Engineering | UC SANTA CRUZ

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

Programs

Education:

- CMPS 107: "Open-Source Programming"
  - TA: Andrew Shewmaker, **Chancellor's Graduate Teaching Fellow** (for syllabus)
  - **Excellence in Teaching Awards** for both TA Andrew and instructor Carlos
  - Speakers: Peter Grehan (TidalScale), Spencer Sevilla (UCSC), H. Peter Anvin, Sage Weil, Jessica Yu (Red Hat)
  - Now a permanent course, next scheduled for Winter 2017, telecasted, webcasted
  - Additional "Advanced Open-Source Programming" planned

Research Projects:

- 6 projects selected for funding
- 1 Graduate Student Researcher (GSR) per project
- Reviewed every 6 months:
  - Fundamental research question
  - Plausible path to successful open-source software project
  - Long-term relevance to companies, university, and society
- Adequate progress, Success: publications, citations, graduation

Incubator Projects:

- 2 projects selected for funding
- 1 Post-doc per project
- Reviewed every 6 months:
  - Out-of-box experience
  - Growing diverse developer community
- Adequate progress, Success: adopted by organization outside university

Baskin Engineering | UC SANTA CRUZ

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

Open Source Experts

- Ensure that the work we do at CROSS is acceptable to the open source software community

Karen Sandler, JD, Executive Director, Software Freedom Conservancy
https://en.wikipedia.org/wiki/Karen_Sandler

Sage Weil, UCSC Alum, Ceph creator, Founder & CTO, Inktank
https://en.wikipedia.org/wiki/Sage_Weil


CROSS
CENTER FOR RESEARCH IN
OPEN SOURCE SOFTWARE

Sponsors


TOSHIBA


SEAGATE


SK hynix memory solutions


Micron

CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

More details

cross.ucsc.edu