www. chameleoncloud.org

“This is Not a Testbed”: How to Build and Operate Experimental Infrastructure

Kate Keahey
University of Chicago / Argonne National Laboratory

keahey@uchicago.edu


Chameleon
1,400+
Unique
projects
1,000+
Papers
published
14,000+
Users
10


Ceci n'est pas une pipe.

WHAT IS A TESTBED?

- **Testbed: "A place where I test things"**
  - Things break – and are supposed to!
- **Scientific Instruments: "Testbed-as-a-Service"**
  - Science-driven "testbed as a service"
  - Adapts infrastructure to solve a specific problem for a specific community
  - Provides **production quality** services
  - Combines operations with **evolution** of features (API-breaking changes may happen)
  - Builds **large** communities working on a **broad range of problems** or specific scientific breakthroughs
- Mainstream Infrastructure

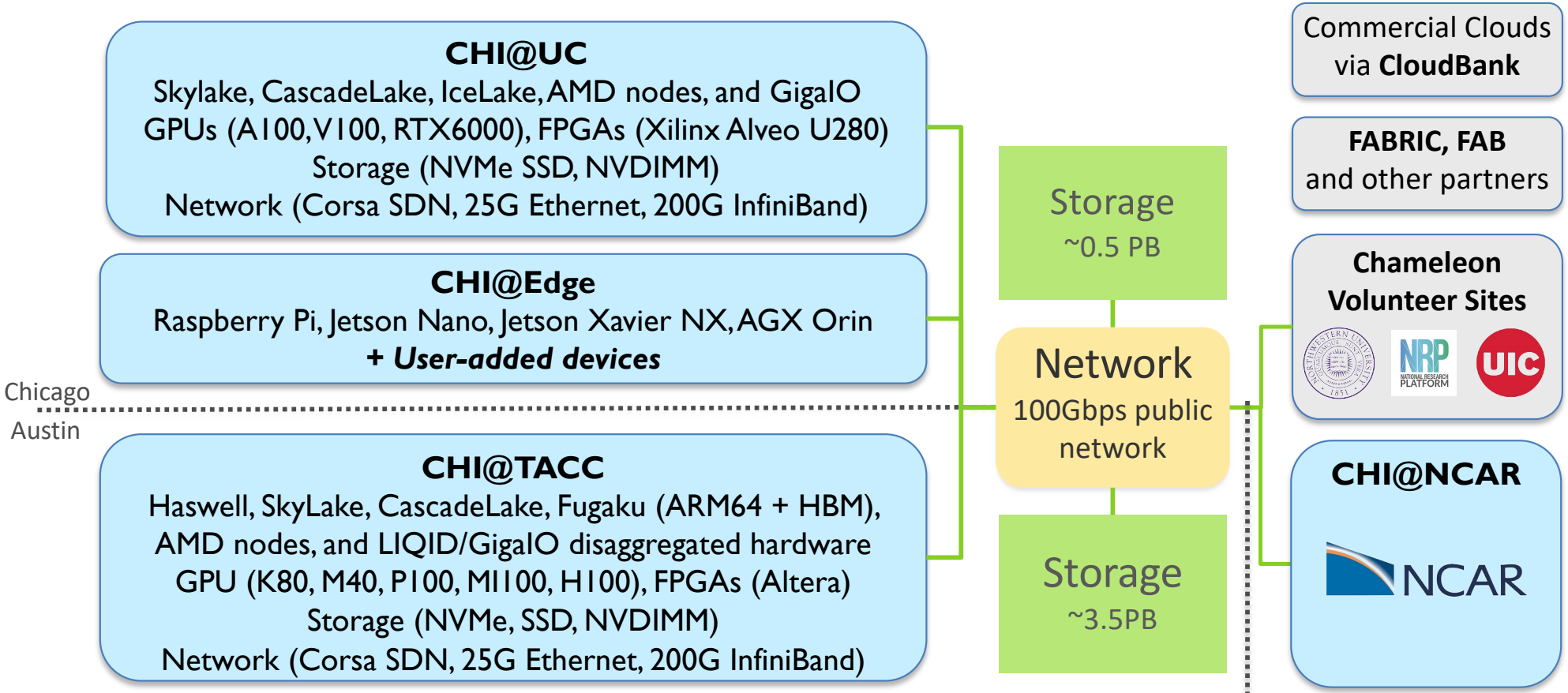
SLICE-BASED RESEARCH INFRASTRUCTURE

STRUCTURAL CONSIDERATIONS

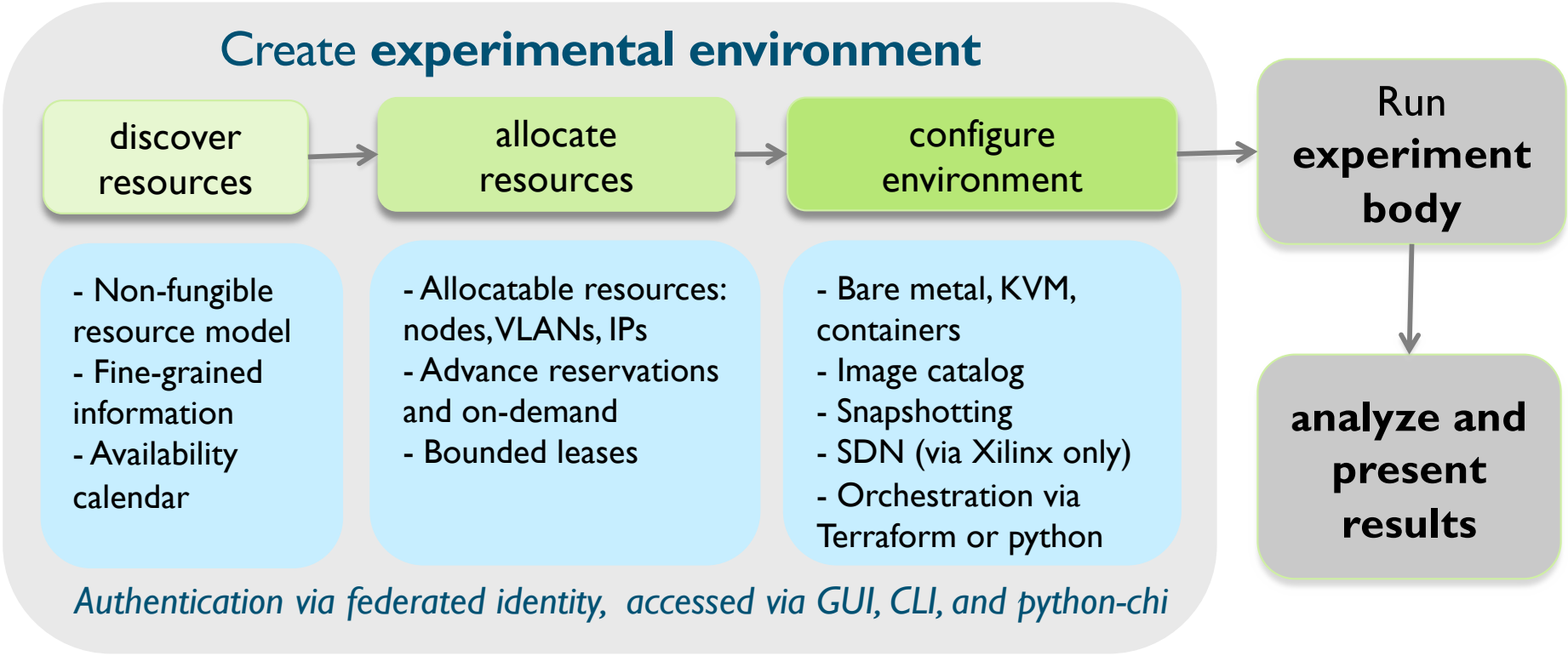
- **Hardware:** support a broad set of experiments
  - Diverse: architectures, accelerators, storage, interconnects, networks, IoT
  - From large to small: scale versus diversity trade-off
- **Capabilities:** deeply reconfigurable
  - Power on/off, custom kernel boot, serial console access, firmware change, etc.
  - A spectrum of reconfigurability options: bare metal, virtualization, containerization
- **Sustainability:** cost-effective
  - Solid base in mainstream open source capabilities
- **Evolution and Adaptability**
  - Capable of evolution: ~50% mainstream + ~50% “special sauce”
  - Packaging: lateral growth with CHI-in-a-Box
  - CHI@Edge and its derivatives
- **Methodology:** supporting and innovating methodology
  - Experiment packaging, management, and sharing
  - Practical reproducibility

openstack
CLOUD SOFTWARE

CHAMELEON HARDWARE



CHAMELEON CAPABILITIES



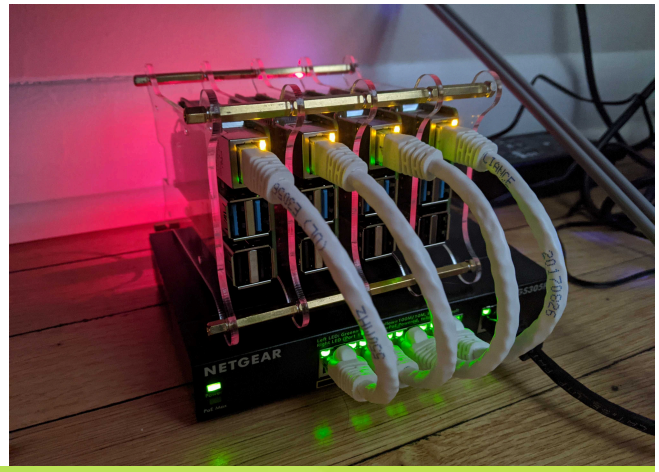
Paper: "Lessons Learned from the Chameleon Testbed", USENIX ATC 2020

CAPABILITIES: CHI@EDGE

A lot like a cloud!
All the features we know and love – but for edge!
"Edge to cloud from one Jupyter notebook."

Not at all like a cloud!
Location, location, location!
IoT: cameras, actuators, SDRs!
Not server-class!
And many other challenges!

- CHI@Edge: all the features you love in CHI, plus:
  - Leverage **Chameleon front-end**
  - Reconfiguration through non-prescriptive **container deployment** via OpenStack interfaces (using K3 under the covers)
  - Support for "standard" **IoT peripherals** (camera, GPIO, serial, etc.) + easy for you to add support for your own peripherals
  - **Bring Your Own Device (BYOD): Mixed ownership** model via an SDK with devices, virtual site, and **restricted sharing** – building on OpenBalena



Paper: "Chameleon@Edge Community Workshop Report", 2021

SUSTAINABILITY: OPENSTACK

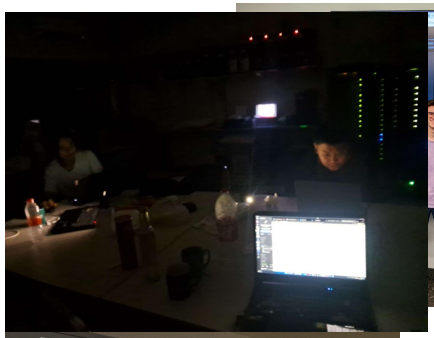
- Using OpenStack + Ironic for bare metal configuration
- Mainstream: 1,000s of deployments, ~3,500+ individual developers contributed, millions of end-users
- The case for mainstream infrastructure
  - Familiar interfaces
  - Transferable skill, workforce development
  - Large contribution community (== able to leverage many features by just upgrading)
  - We can contribute multiplying our broader impacts
  - Ties us to mainstream software ecosystem
- "Special sauce"
  - Federated identity, bare metal snapshotting, resource discovery and availability, and others
  - Advance reservations for bare metal – contributed to OpenStack

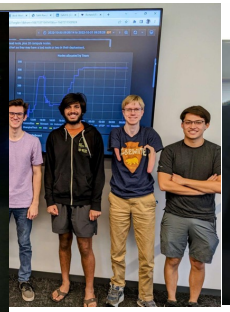
SUSTAINABILITY: CHI-IN-A-BOX

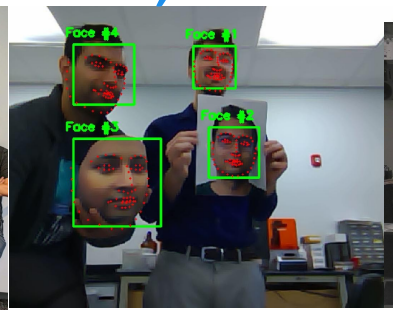
- CHI-in-a-box: packaging of CHameleon Infrastructure (CHI)
  - Internal packaging of a commodity-based testbed
  - Packages the system as well as the operations model
  - Hub and spoke management, version-controlled site configuration management as code, containerization, monitoring, detection, and remediation tools
  - Support for Bring Your Own Device (BYOD) model: Doni allows administrators to dynamically enroll resources, define availability windows, and streamline operations
- Deployment
  - Deployed volunteer sites (overtime): IIT, NCAR, Northwestern, Purdue, and UIC
  - Independent testbed: ARA
  - In conversation: SDSC, OCT/U Mass, FIU, ORNL, KTH (edge/wireless only), NUS, and others

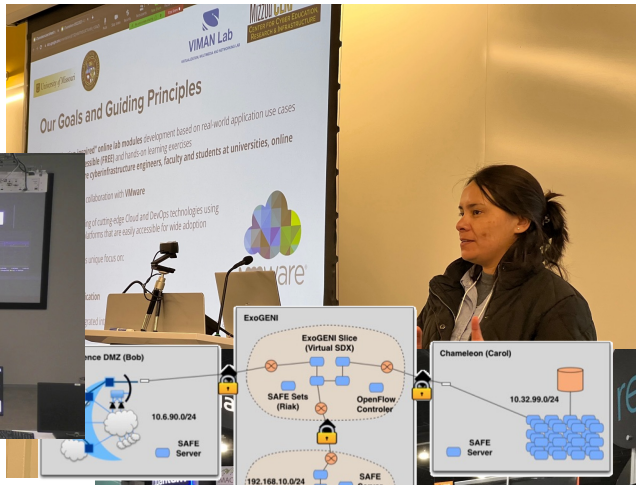
Paper:"CHI-in-a-Box: Reducing Operational Costs of Research Testbeds ", PEARC'22

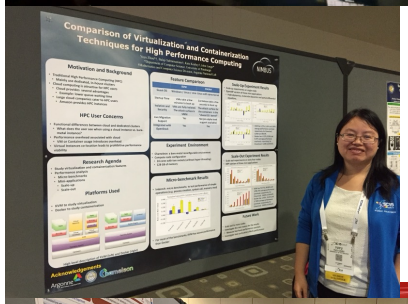
NOT JUST A TESTBED, A COMMUNITY



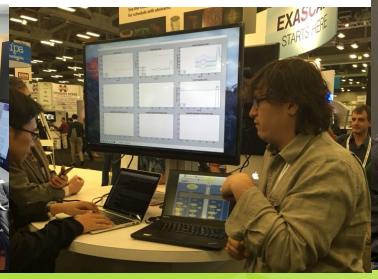


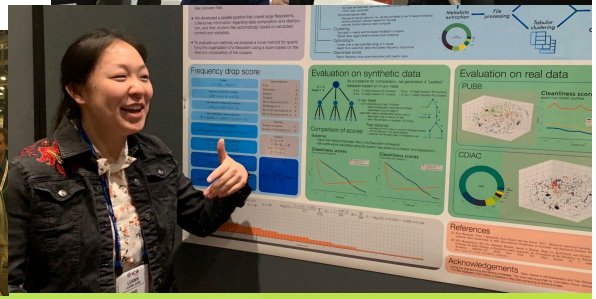






Supporting research projects in architecture, operating systems design, virtualization, power management, real-time analysis, security, storage systems, databases, networking, machine learning, neural networks, data science, and many others.





Chameleon www.chameleoncloud.org

Check out user experiment stories on our blog:
https://www.chameleoncloud.org/blog/category/user-experiments/

RESEARCH: AI-DRIVEN ENERGY OPTIMIZATION

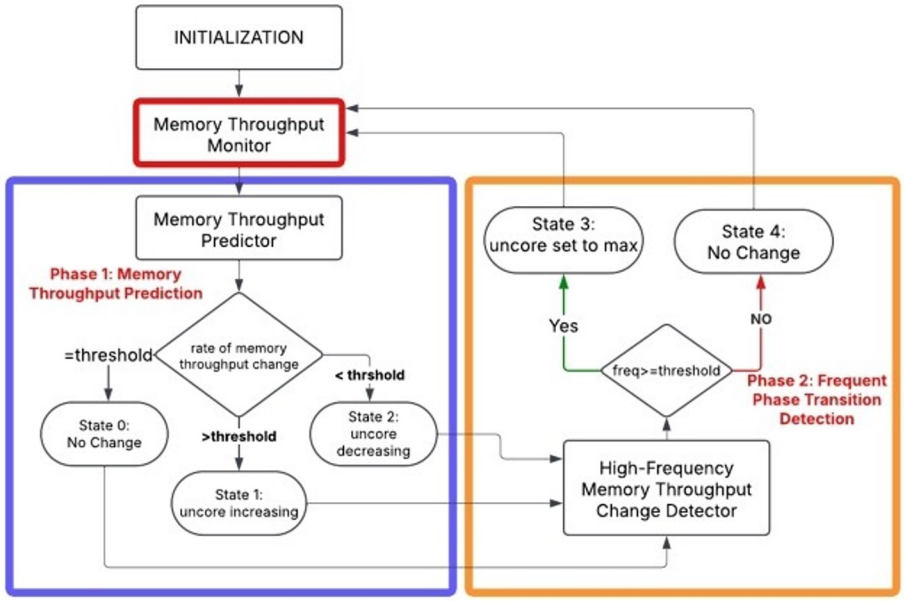
Research Question: Can AI models predict and optimize power consumption of GPU workloads to reduce energy waste in heterogeneous computing systems?

What They Did:

Developed system using ML to adjust CPU power settings based on real-time memory usage patterns
Monitored when applications need high vs. low memory bandwidth to minimize wasted power
Detected rapid workload changes to maintain performance during critical periods

How They Used Chameleon:

Conducted experiments on nodes with Intel Xeon Platinum 8380 CPUs and NVIDIA A100 GPUs
Tested across GPU benchmarks, HPC applications, and ML training workloads (UNet, ResNet50, BERT)
Achieved ~27% energy savings with <5% performance loss



Paper: Zheng et al. "Minimizing Power Waste in Heterogeneous Computing via Adaptive Uncore Scaling", SC '25

EDUCATION: TEACHING OPERATIONAL ML

Objective: develop skills to design, build, and operate ML systems

- Topics: cloud computing, DevOps for ML systems, model training at scale, model serving, monitoring and evaluation, data systems, safeguarding ML systems
- Student projects organized around building a startup based on ML systems

Course details

- Enrollment: 191 students
- Grading: 60% weekly labs, 40% group project
- Project: Students worked in groups of 3-4 to design and implement a large-scale ML system

Project: Fraida Fund, NYU

Platform tradeoffs

	User has control over compute + storage + network + systems?	Limits $ risk?	Like a "standard" cloud?
Conventional HPC	✗	✓	✗
Commercial clouds aws	✓	✗	✓
Other research testbeds FABRIC	✓	✓	✗
Chameleon Cloud	✓	✓	✓

Chameleon resources used for lab assignments

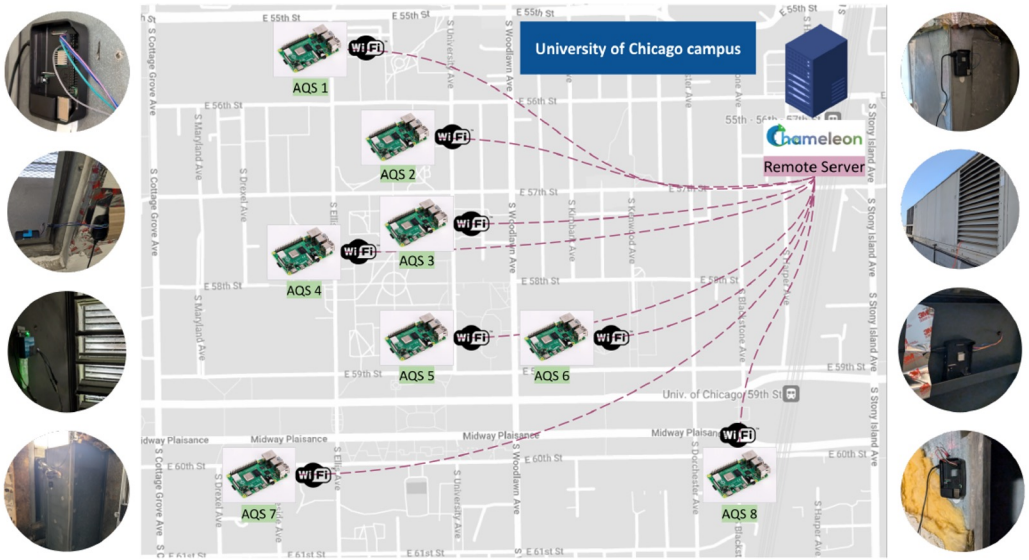
Compute	Basic VM compute instances (`m1.small`, `m1.medium`, `m1.large`) Single-GPU instances (`compute_liqid`, `compute_gigaio`, `gpu_rtx6000`) Multi-GPU instances (`gpu_a100`, `gpu_v100`, `gpu_p100`, `gpu_mi100`) Edge devices (`raspberrypi5`) (BYOD)
Network	Floating IP addresses Virtual private networks + routers Security groups to permit ports on which services are running
Storage	Block storage volumes Object storage
Interface	Browser-based GUI (OpenStack Horizon GUI) CLI (`openstack` CLI) via Chameleon-hosted Jupyter environment Python API (`python-chi`, OpenStack Python API) via hosted Jupyter Terraform via OpenStack provider

RESEARCH EXAMPLE: FEDERATED LEARNING

Compare simulation, emulation, and real-world deployments for **Federated Learning**

Deployed **8 Raspberry Pis with air quality sensors** on UChicago campus

Simulating FL training on a single compute node can **accurately reproduce model performance metrics** (accuracy), but presents **limitations for reproducing system metrics** (training time, CPU usage, and communication latency)

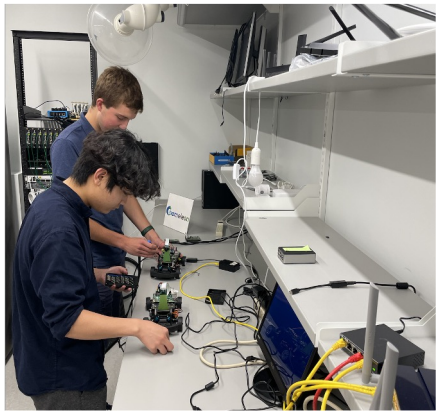


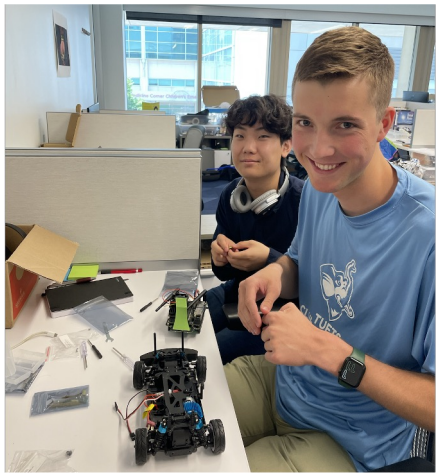
Paper:"On Reproducibility Challenges of Federated Learning: Investigating the Gap between Simulation, Emulation and Real-World Deployments", CCGrid'25

AUTOLEARN

- Chameleon notebooks based on the DonkeyCar package
- Students learn in three stages:
  - Data collection – actual/simulator – edge to cloud
  - Model training in the cloud
  - Verification via autonomous driving – actual/simulator – edge to cloud
- Supports different emphasis in teaching
  - Introduction to engineering might emphasize driving the actual car
  - Machine learning focus might use the simulator
- Individual exploration:
  - E.g., digital twin combining simulator and experimental driving

Paper: "AutoLearn: Learning in the Edge to Cloud Continuum", EduHPC'23





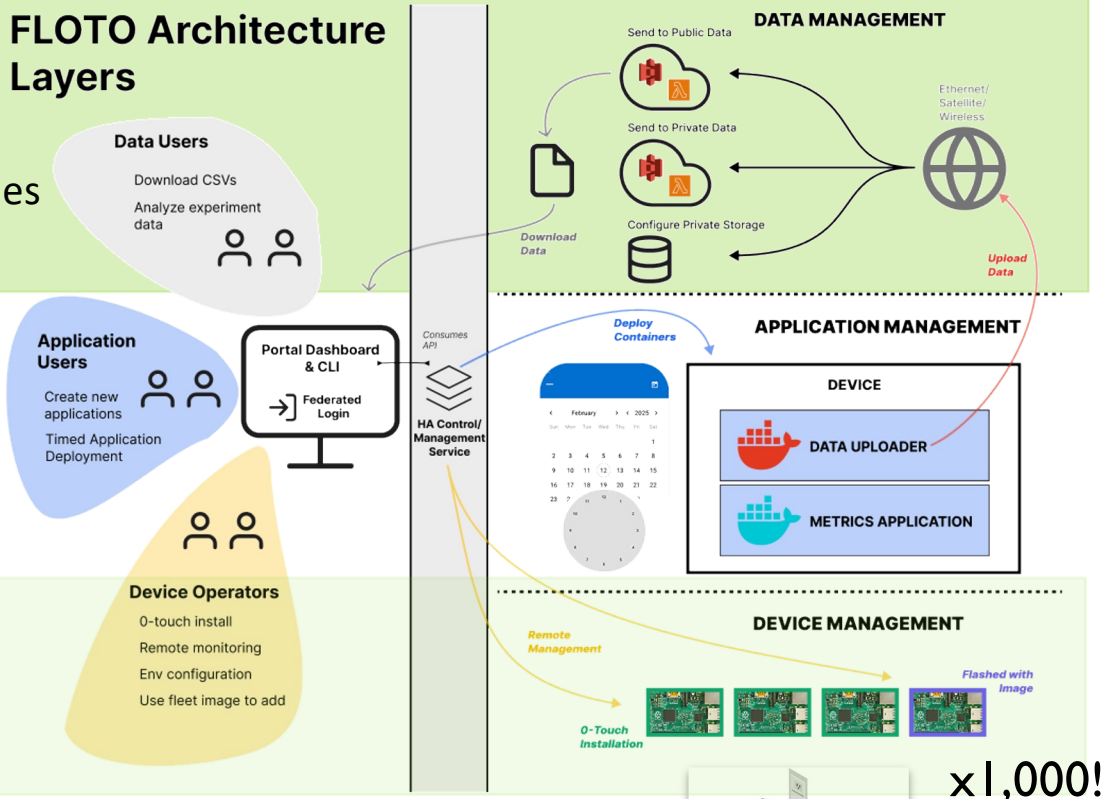
REU 2023 students working on hardware setup for autonomous vehicles, packaging a curriculum developed by Rick Anderson, Rutgers

ADAPTING INFRASTRUCTURE: THE FLOTO PROJECT CASE STUDY

- Why broadband monitoring?
  - Technical questions: what happens in conditions of oversubscription?
  - Policy questions: can we characterize the “digital divide” in our society?
  - Modeling questions: what assumptions about broadband are realistic?
- Measuring broadband – different approaches/applications depending on context, objective, use case, etc.
  - Netrics: open-source library of standard network diagnostic tools (ndt7, speedtest, ping, traceroute, etc.) for continuous, longitudinal network measurement
  - Others: e.g., residential versus rural broadband and other use cases
- **Can we use CHI@Edge as a large observatory instrument for broadband monitoring?**
- **Approach:** connect a “measurement box” to the router and run tests
- Collaboration with Nick Feamster & his UChicago team

IBIS: A SENSING SUPERCOMPUTER

- Data users
  - Sharing versus privacy trade-off
  - Established community data pipelines versus new sharing methods
- Application users
  - Applications composed of several functions (pods)
  - Application configuration
- Device operators (BYOD layer)
  - Ease of use vs control trade-off
  - User operator vs centralized ops



x1,000!

INSTRUMENT ADAPTABILITY

What knobs can I turn on this instrument?

- Deployment scope: deploy the devices in a different area
- Application: adapting "sensing abilities" programmatically
- Hardware: combine devices with different IoT gadgets (e.g., GPS)
- Data aggregation: different methods for different applications
- Data: ask different questions of the data

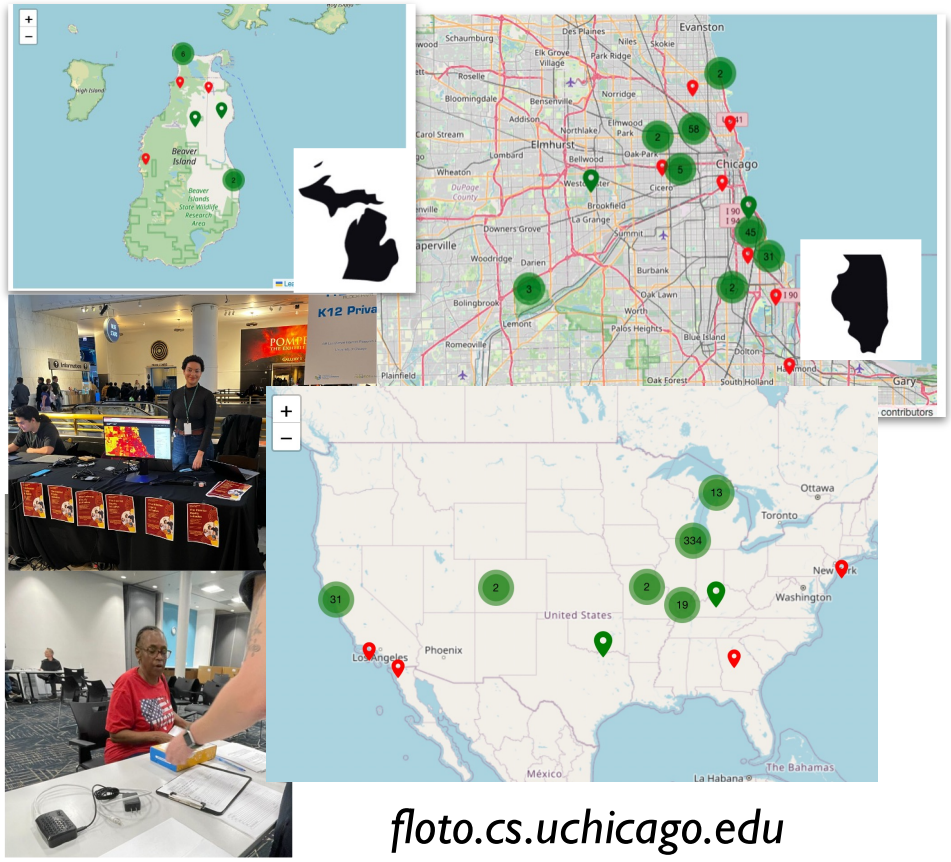
FLOTO: DEPLOY DEVICES IN DIFFERENT AREAS

~500 devices deployed across multiple states
Notable deployments:

- Chicago (180+ devices)
- Milwaukee (200+ devices)
- Marion County, IL; Beaver Island, MI -- and others

As a distributed community, we rely on trust and deep partnerships to bring infrastructure where it is needed most

- Building trust with communities
- Managing devices remotely (with many participants)
- Coordinating large-scale distribution



floto.cs.uchicago.edu

FLOTO: RUN A DIFFERENT APPLICATION

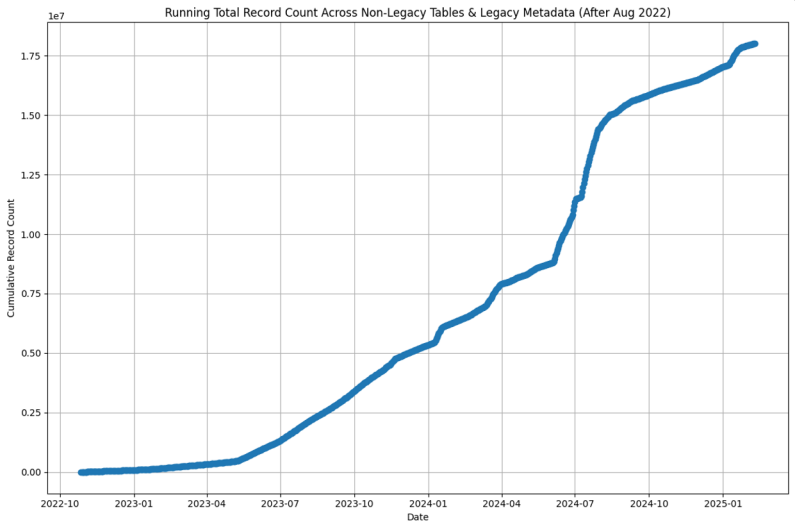
Applications Deployed on FLOTO:

Each application provides different methods for broadband measurement depending on research interest

- **Netrics**: Broadband performance measurements to study access networks
- **RADAR Toolkit**: QoE measurements for telehealth applications
- **NetUnicorn**: Data pipeline experiments
- **Georgia Tech**: IPv6 Performance Studies
- **M-Lab**: Measurement Swiss Army Knife (MSAK) integration
- **ARA**: Monitoring 5G wireless performance in rural areas

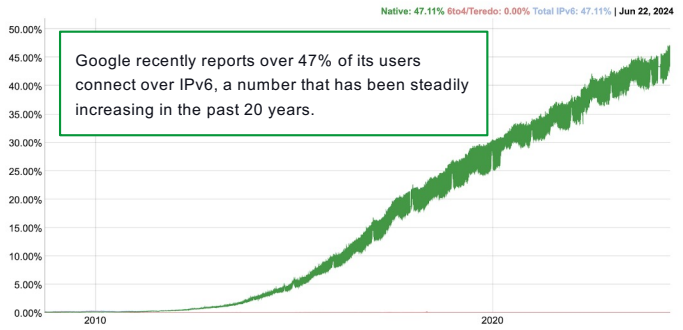
FLOTO: MINE THE DATASET

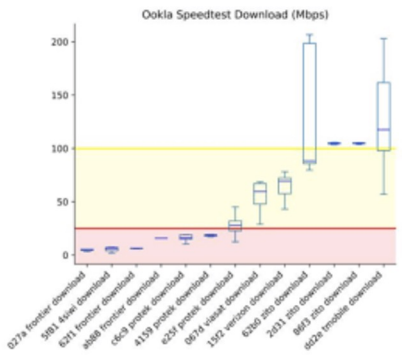
- ~36M million measurements collected since Oct. 2022
- What Measurements? Time series speed tests, latency, DNS performance, network paths on fixed connection (no WiFi bias)
- Spans 17 different network providers
- Multiple access technologies (fiber, cable, satellite, fixed wireless)
- Data is publicly available via project website

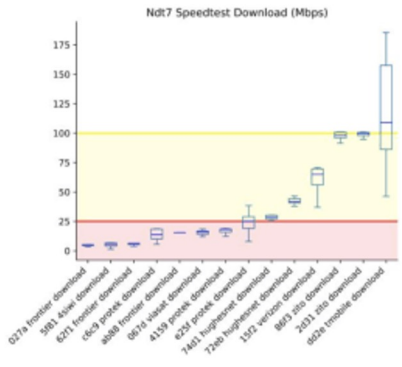


FLOTO: CASE STUDIES

- Computer Science questions: IPv4 versus IPv6
  - Objective: Understand how Internet speed varies between IPv4 and IPv6’
  - Method: sequential speed tests comparing IPv4 and IPv6 results under similar conditions
  - Early Findings: IPv4 and IPv6 speeds degrade differently under various conditions, influenced by the ISP (SIGMOD paper)
- Policy questions: Marion County
  - Objective: Improve internet infrastructure and performance in Marion County, Illinois
  - Method: Deploy FLOTO devices to collect and analyze broadband performance data
  - Finding: 32% of sampled households below the federal threshold -- data used to support grant applications for fiber broadband expansion

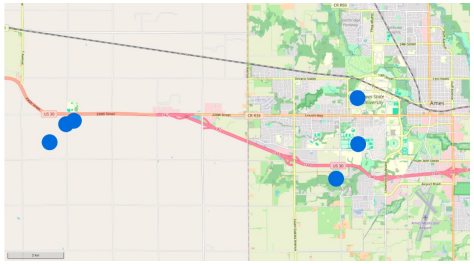


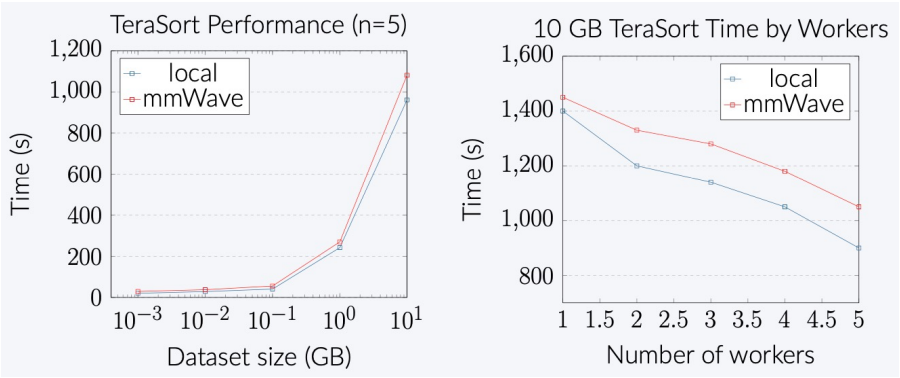




MEASURING RURAL WIRELESS

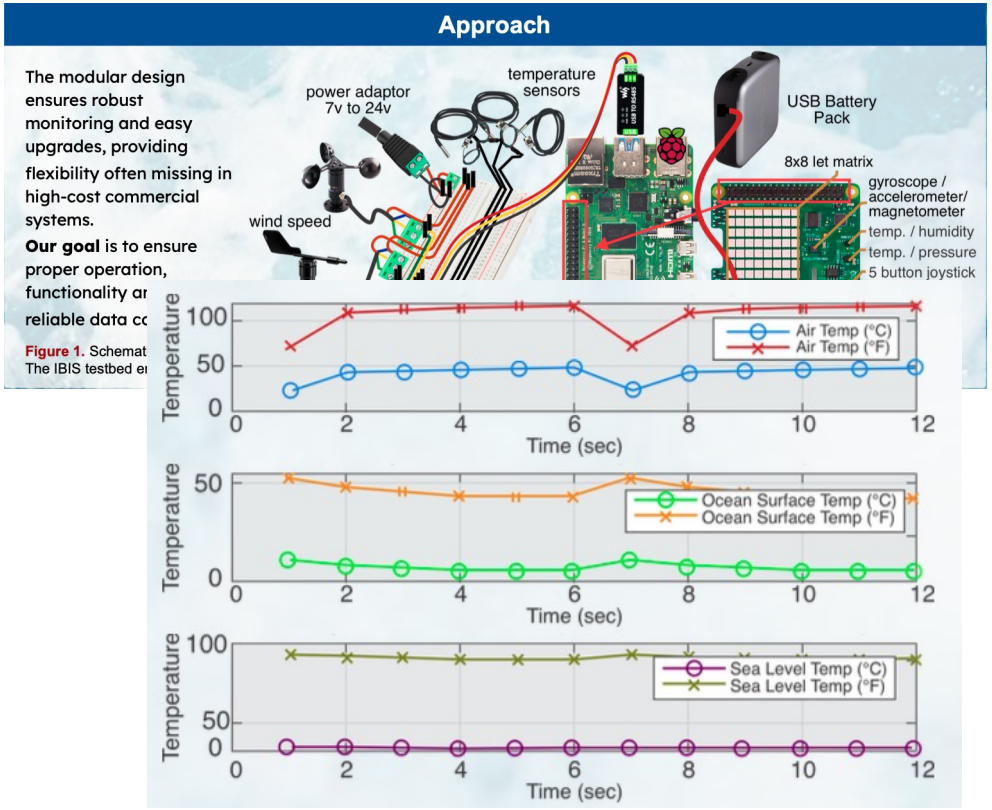
- Collaboration with ARA project
- Assessing the quality of rural 5G networks
  - Measuring device to device latency
  - Clock synchronization
  - Comparing over different network fabrics
- Deployed 6 Raspberry Pi devices with 5G connectivity in rural Iowa
- Latency measurements: GPS-based time synchronization for precise measurements (4000x more precise than NTP over 5G)
- Tested using Hadoop
- Hey presto: 5G networks can support distributed computing with performance comparable to wired connections!





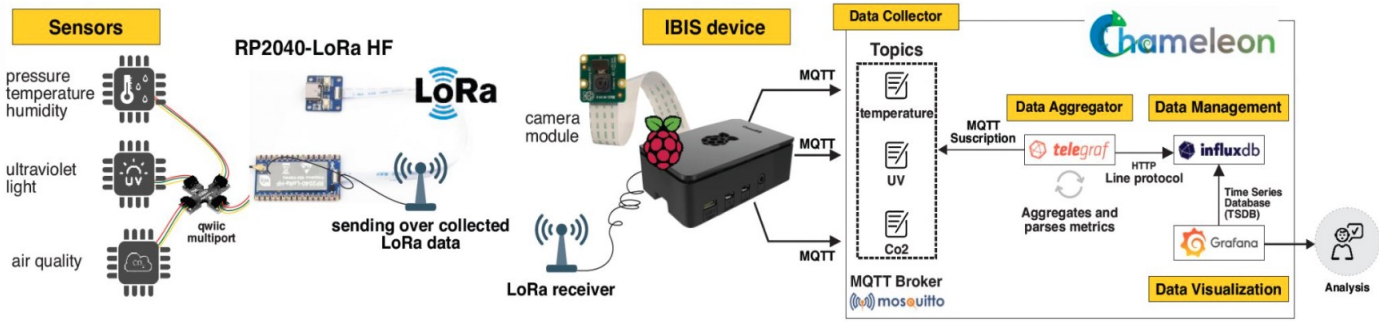
SENSOR STATIONS FOR MARINE AND COASTAL ECOSYSTEMS

- Smart buoy system: sensor stations for oceanic data collection (water quality, water movement, water levels, etc.)
- Collaboration with FIU
- Integrated multiple environmental sensors with IBIS infrastructure
- Demo deployment with real and simulated data
- Implemented cloud-based data visualization system
- Collaboration with FIU



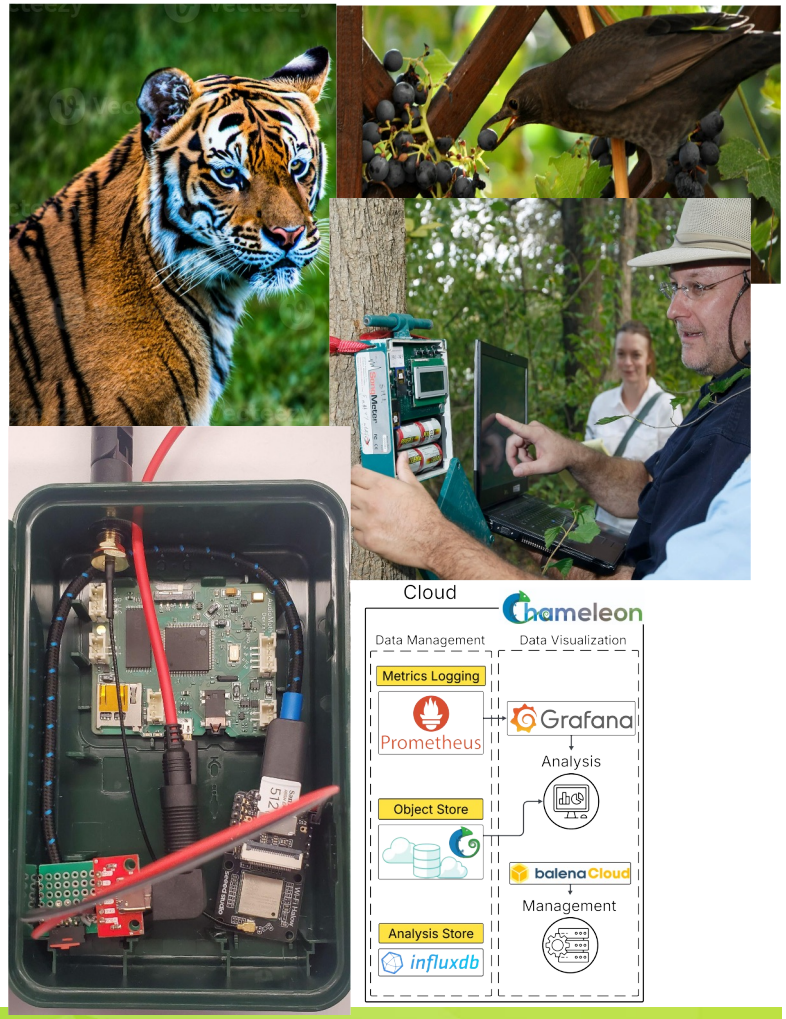
NCAR WEATHER SENSING STATIONS

- openIoTwx: NCAR 3D printed weather stations
- Richer continuum: IBIS SBCs connecting to openIoTws via LoRa
  - Exploring power (4x factor), connectivity (cellular vs aggregation via LoRa), sensing (additional camera sensors), and processing (to e.g., reduce size of data) trade-offs
- Future challenges
  - Image-based weather prediction methods, scaling up to create dense, high-resolution weather monitoring networks, and assessing long-term reliability in diverse outdoor environments



SOUNDSCAPING

- Using acoustics for biodiversity conservation: tracking wildlife, protecting crops
- Scaling challenge
  - Expensive hardware (~$1,000 per device)
  - Requires manual data collection and servicing
- How can we
  - Reliably stream and analyze audio in **real-time**
  - From **thousands of Listeners**, not dozens
  - While minimizing hardware and operating **costs** for years-long studies
  - In an environments integrating deployment, visualization, storage, and management
- Architecture: custom low cost/power Listeners and Aggregators combine needs-based sensing with network access with an integrated data analytics framework



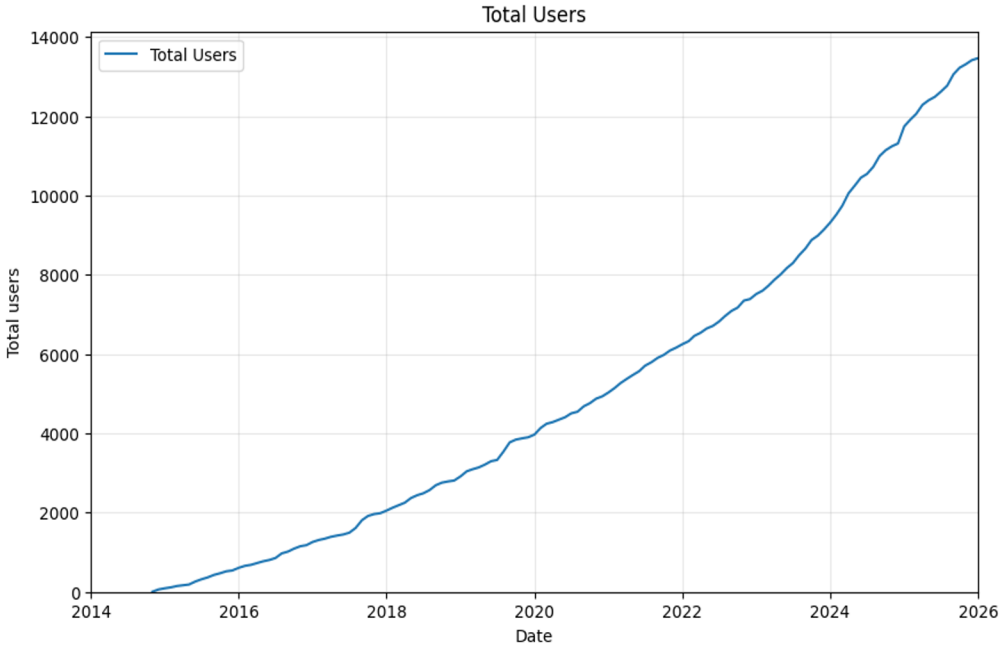
Hudson Reynolds, Boston University
Alex Tuecke, Worcester Polytechnic Institute

HOW DOES THE SYSTEM GROW?

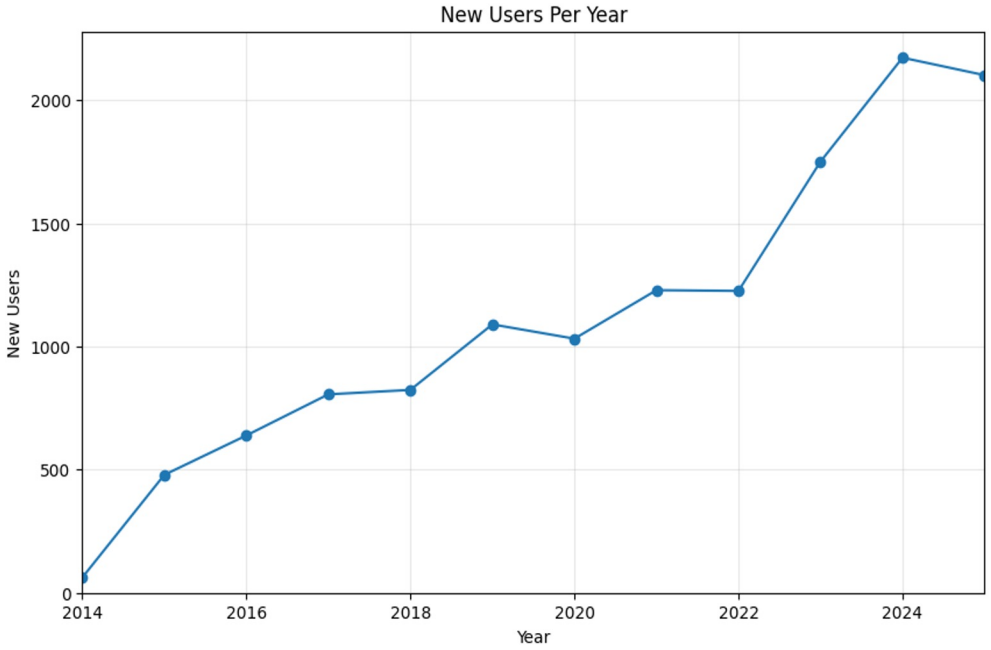
(AND WHAT MEASURES ARE REASONABLE TO USE TO ASSESS THIS GROWTH?)

Paper: "Practical Evaluation Methods for Scientific Instruments", PEARC 2026

HOW DOES THE SYSTEM GROW? (USERS)

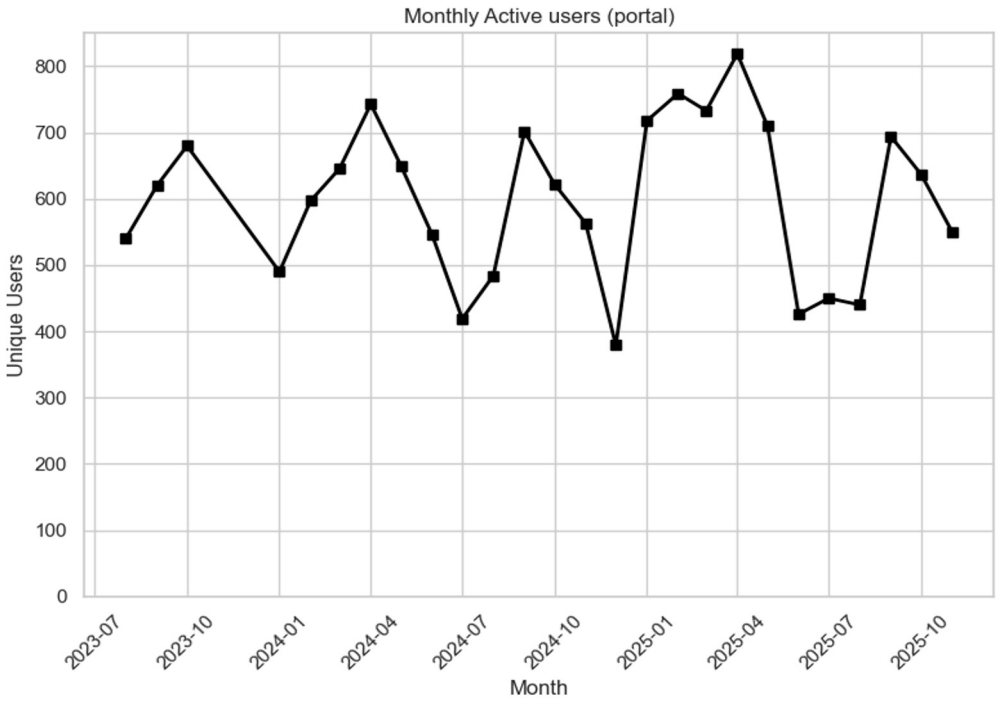


Total users per year

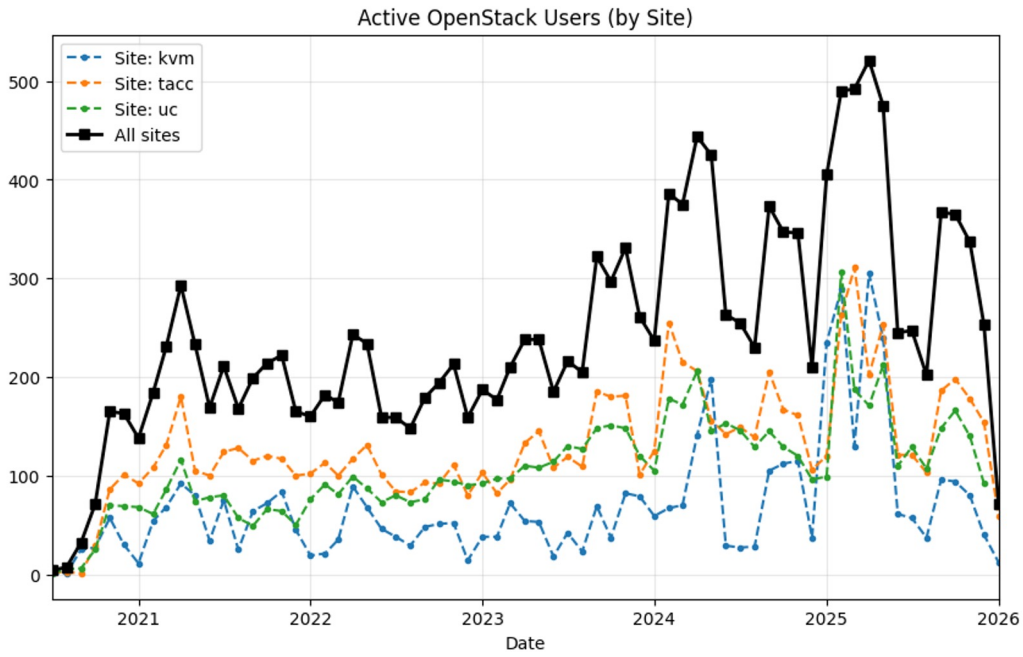


New users per year

HOW DO USERS USE THE SYSTEM?

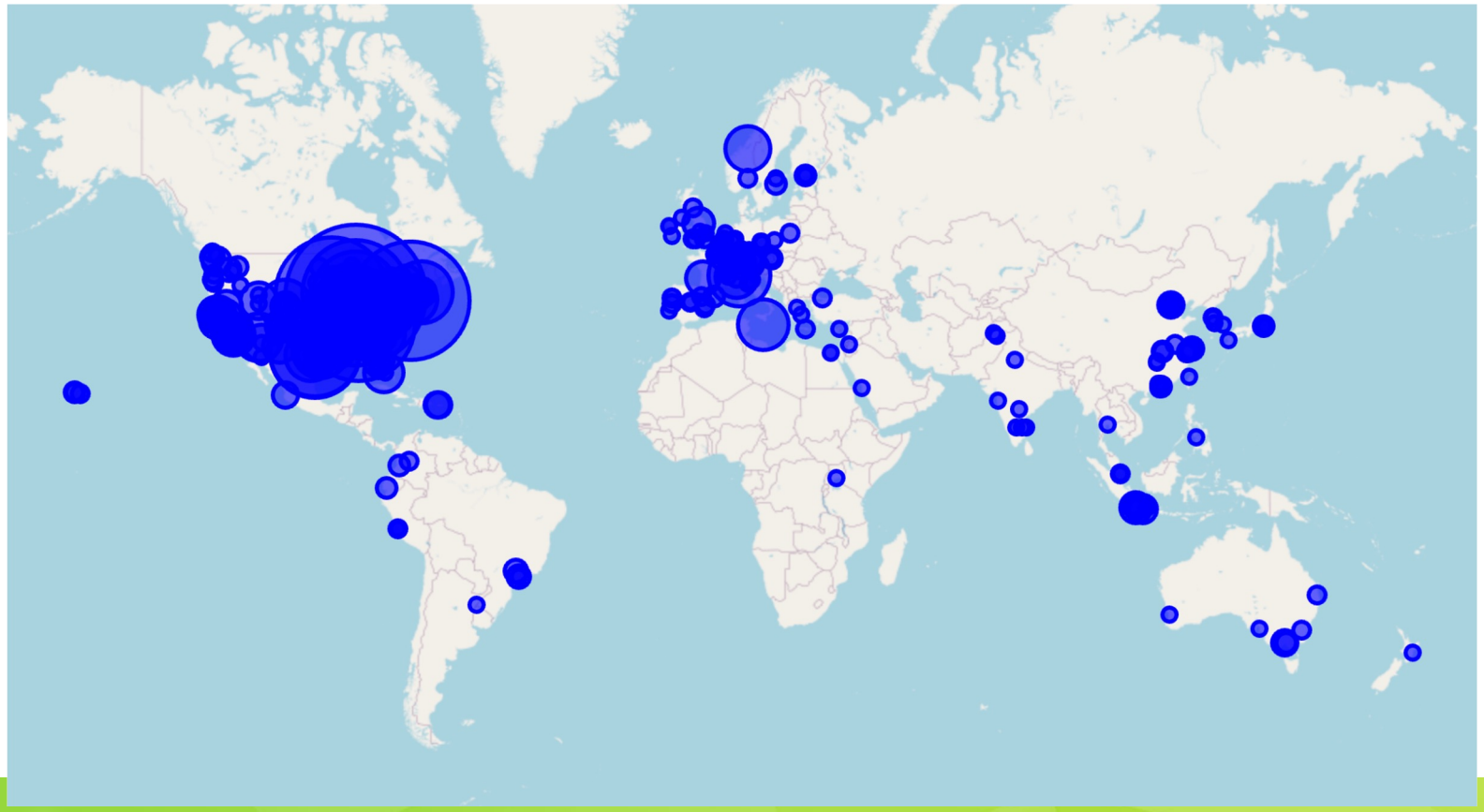


Unique users in the portal per month



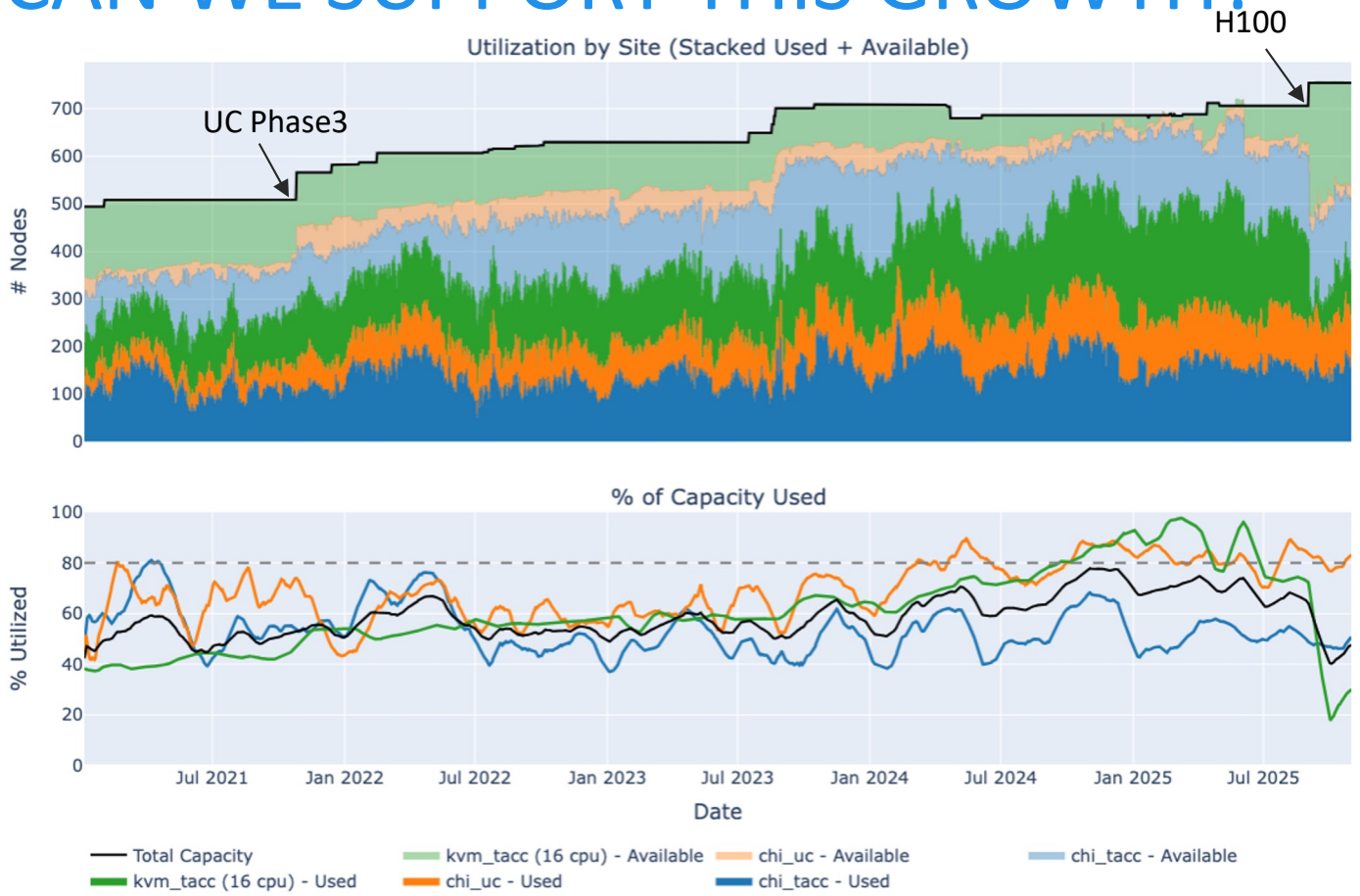
Unique users per OpenStack site, per month

INSTITUTIONS ACROSS THE WORLD



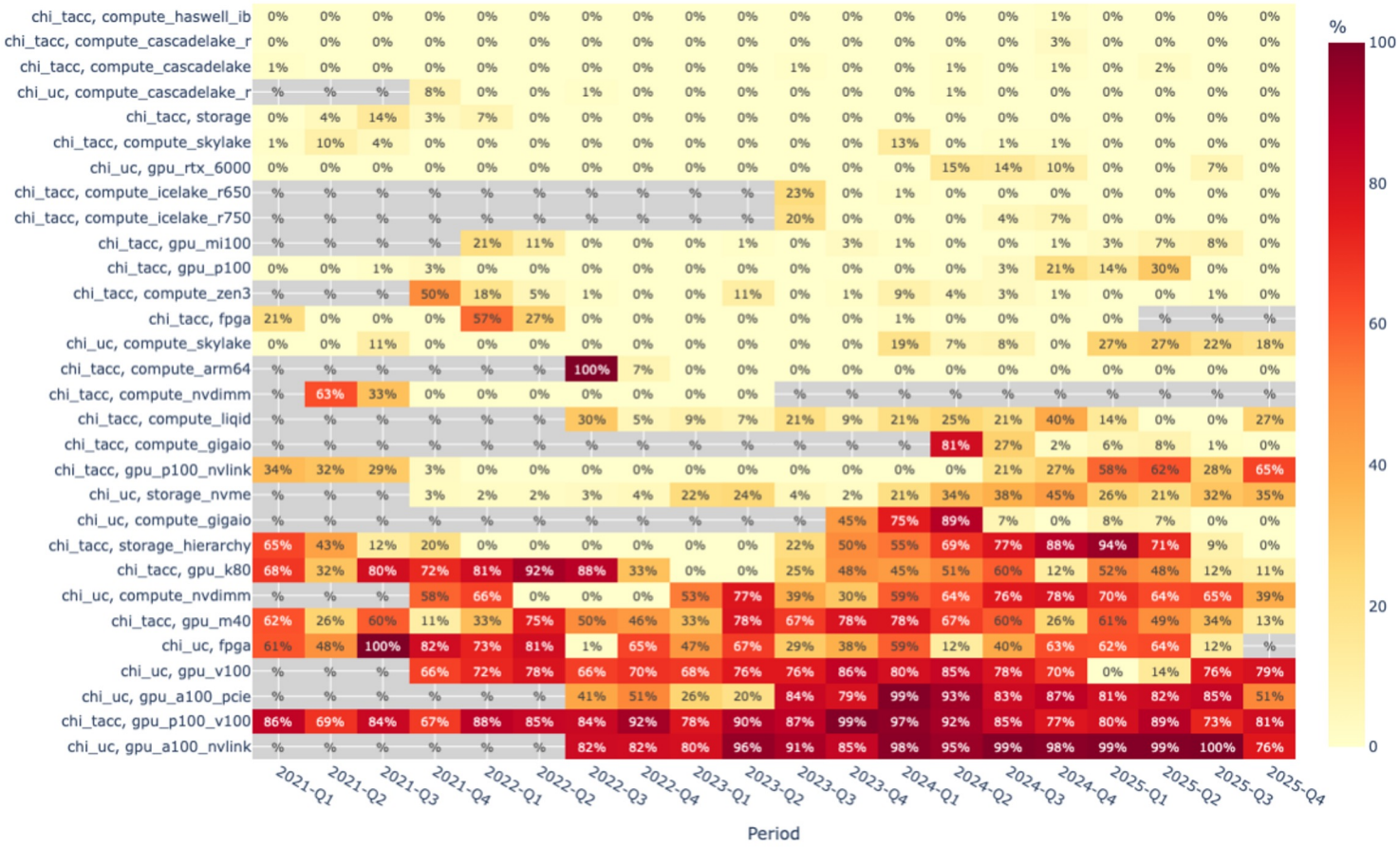
HOW WELL CAN WE SUPPORT THIS GROWTH?

- Hitting the highs: CHI@UC and KVM@TACC neared 100% in 2025!
- Note: VM capacity is translated 16 vCPUs to a node



POPULARITY CONTEST!

- 80% -> 80% of the time, no resources of this type are available
- capacity constrained:
  - GPUs
  - High Memory
  - Fast Storage
- Specialized hardware has intermittent demand (FPGA, GigaIO)

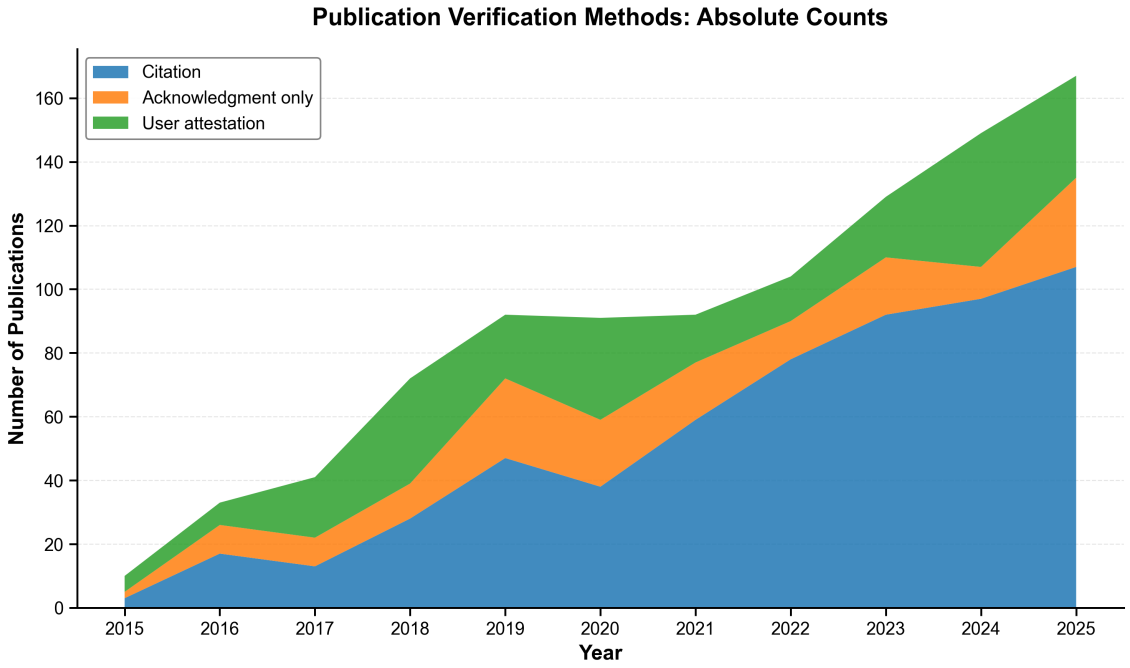


WHAT IS THE COLLECTIVE SCIENTIFIC OUTPUT OF THE SYSTEM?

- Challenges: attribution, identification, de-duplication, inclusion
- Attribution methods
  - Citation of Chameleon papers
  - Written acknowledgment in text
  - User attestation post-publication
- Identification methods
  - User Reporting via allocation renewals or web form (explicit attestation required)
  - Algorithmic Scraping via scopus, semantic scholar, and Google scholar
- De-duplication:
  - Manual review using similarity matching (difflib SequenceMatcher, 0.7 threshold)
  - API identifier checking to prevent re-importing from same source
- Including only papers using Chameleon to produce experiment results

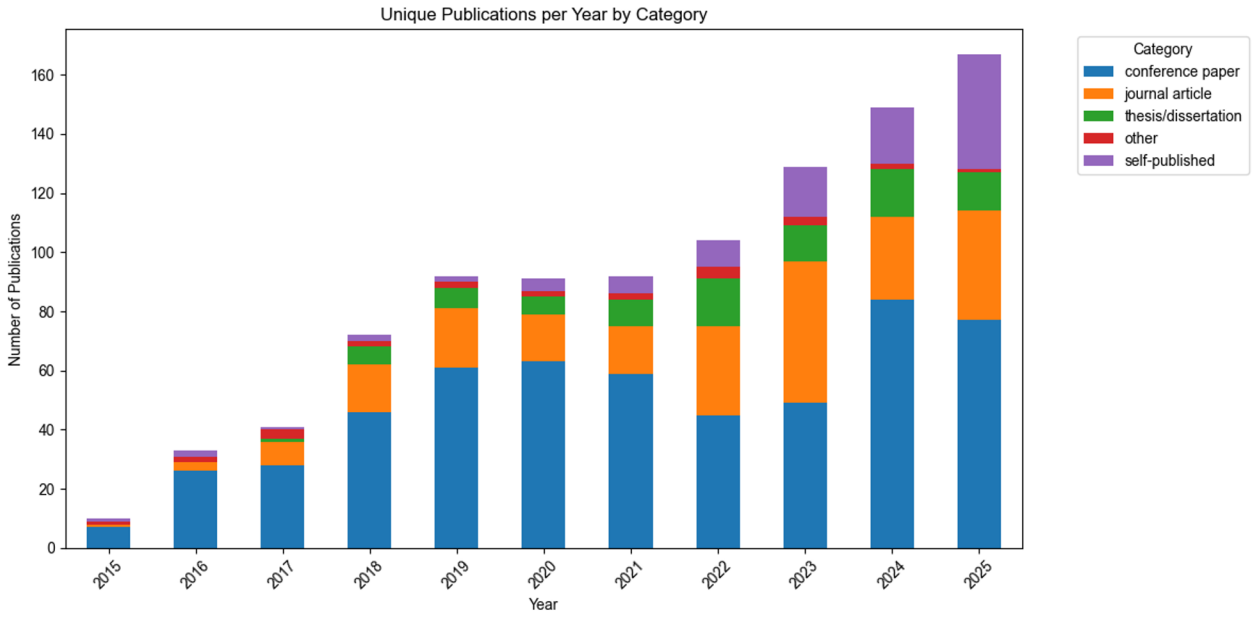
PUBLICATIONS BY IDENTIFICATION

- Total publications grown every year (except 2021) to 982 total through 2025 (currently 1,027)
- How are they identified?
  - ~62% by citation
  - ~18% by acknowledgement
  - ~20% by user attestation



PUBLICATION TYPE BY YEAR

- 982 unique publications produced by Chameleon users (2015-2025)
- True lower bound
- Main type breakdown
  - 55% - conferences
  - 23% - journals
  - 8% - theses/dissertations
- “Other” category includes books, patents, posters, software/data, and tutorials
- 105 total publications are preprints or presentations



PUBLICATION VENUES

Users publish “everywhere”: breakdown of venues for the top 100 most cited papers produced using the testbed:

USENIX - 7 papers (FAST, HotEdge, OpML, OSDI/ATC, NSDI)
IEEE - 30 papers (CLOUD, CLUSTER, DSN, e-Science, IC2E, ICCD, ICIP, ICMLA, ICPP, ICSTS, Internet Computing, IPDPS, MILCOM, SEC, SP, and various IEEE-T journals)
ACM - 26 papers (APLOS, Computing Surveys, CoNEXT, FSE, HPDC, ICPP, ICSE, IJCAI, ISCA, Management of Data, Middleware, PEARC, **SIGCOMM**, SIGENERGY, SIGMOD, SoCC)
Joint ACM/IEEE - 5 papers (CANOPIE-HPC, SC, WORKS)
Other CS journals/conferences - 24 papers (AAAI, ACL, CIDR, Computer Networks, EuroPar, EuroSys, Future Generation, Journal of Parallel and Distributed Computing, MLSys, NeurIPS, PNAS, R Journal, VLDB)

Highlighted venues appeared more than once; ACM SIGCOMM appeared the most (4 times)!

METHODOLOGY: BUILDING AND SHARING EXPERIMENTS

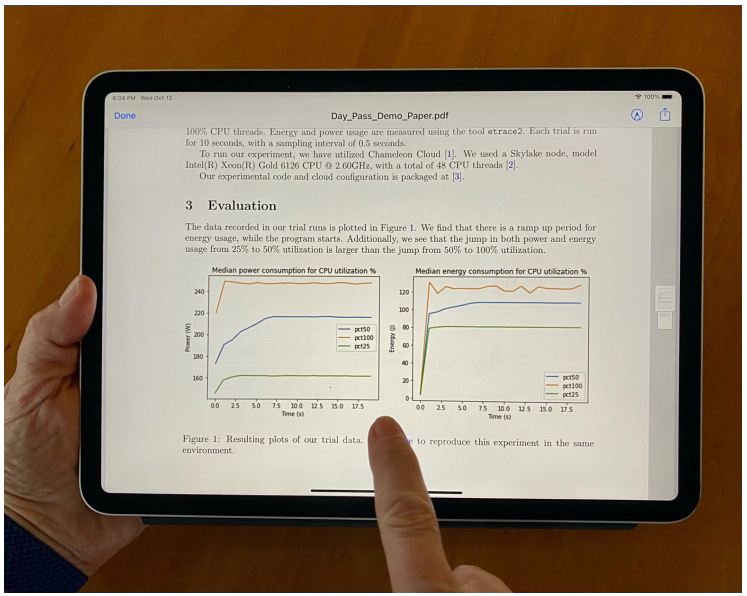
- **Open slice-based platforms:** essential for sharing
  - Hardware accessible to all
  - Non-fungible resources
- **Experimental environment setup:** hardest problem in CS experimentation
  - A user has to **create a custom environment** – and might as well **save it** (snapshot)
  - **Programmatic infrastructure interface and tools** (Heat, Terraform, python-chi jupyter notebooks)
  - Treasure trove: thousands of images, orchestration templates, digital artifacts of various kinds – all ready for sharing!
- How do we harness this treasure trove to simplify and share research?

Paper: "The Silver Lining", IEEE Internet Computing 2020

METHODOLOGY:
BUILDING AND REPRODUCING EXPERIMENTS

Practical reproducibility == feasible enough to be a mainstream method of scientific exploration

- Can digital experiments be as sharable as papers are today?
- Is there a library I can go to and find experiments to play with?
- Can I simply integrate somebody's model into my research instead of reinventing the wheel and get to a new result faster?
- Can I discover something new through playing with somebody else's experiment?
- Can I develop exercises for my class based on most recent research results?

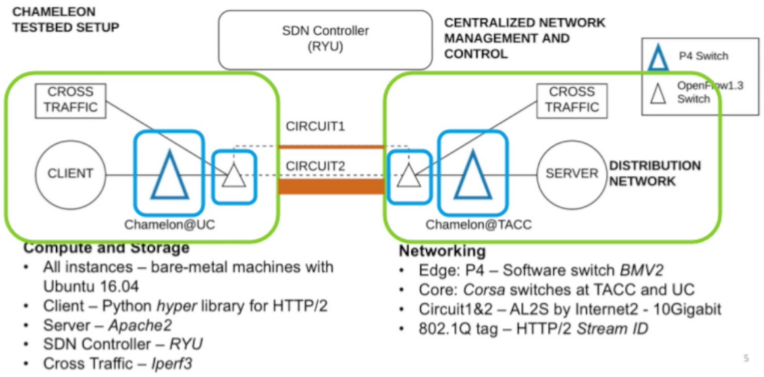


https://repeto.cs.uchicago.edu

PACKAGING EXPERIMENTAL ENVIRONMENTS

- **Declarative methods**
  - Heat, Terraform, and other mainstream orchestration tools
  - Hard to introspect and could be tricky for reproducibility
- **Imperative methods**
  - CLI, python-chi and/or scripts
  - Potentially via Jupyter integration
  - Can be re-played incrementally, troubleshooting and making changes as you go
- **The user chooses the method**

Package this!



Complex Experimental containers via programmable interfaces

Paper: "A Case for Integrating Experimental Containers with Notebooks", CloudCom 2019

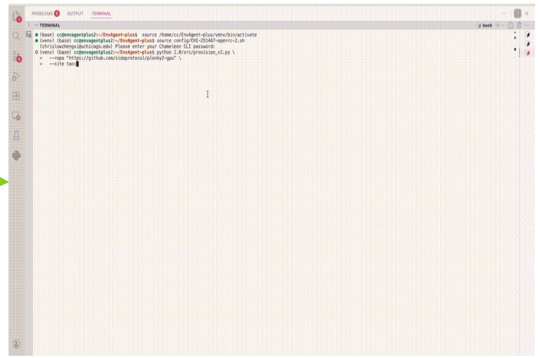
YOUR AI-DRIVEN EXPERIMENT ASSISTANT

Our implementation targets high-throughput cryptographic workloads by offloading three primary computations to the GPU: Fast Fourier Transforms (FFTs), Merkle tree construction, and polynomial operations.

To ensure reliable performance and reproducibility of our results, all experiments were conducted on a single-node GPU system with the following minimum hardware configuration: an 8-core CPU, 16GB of system RAM, and an NVIDIA GeForce RTX 2080 Ti GPU equipped with 12GB of GPU memory.

The software stack requires NVIDIA CUDA version 12.0 or newer, along with compatible GPU drivers.

- Read the requirements
- Find and reserve Chameleon resources
- Boot the instance, attach a floating IP, install software dependencies
- Run validation tests

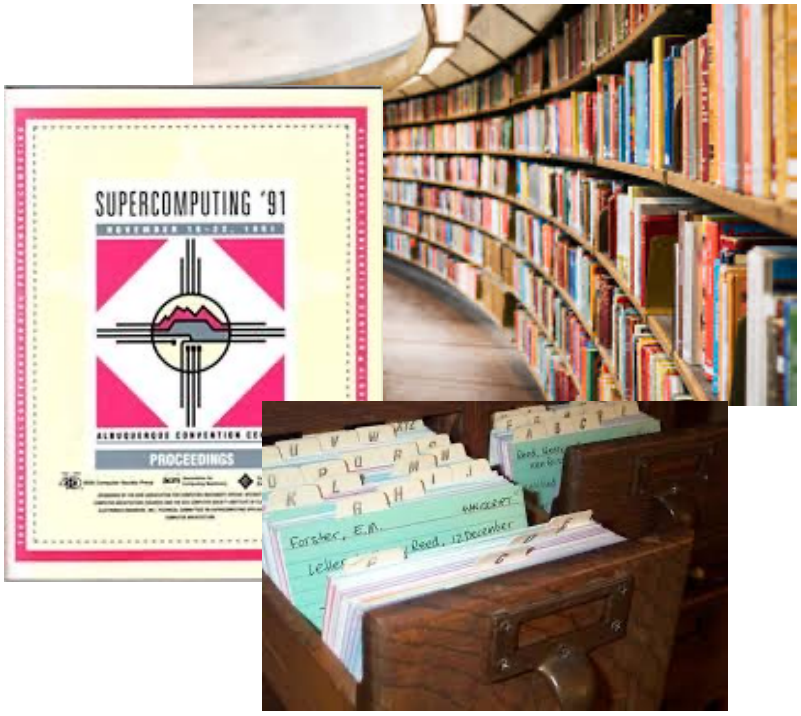


Start at ai.chameleoncloud.org

(documentation chatbot only for now)

SHARING, FINDING, AND REPLAYING EXPERIMENTS

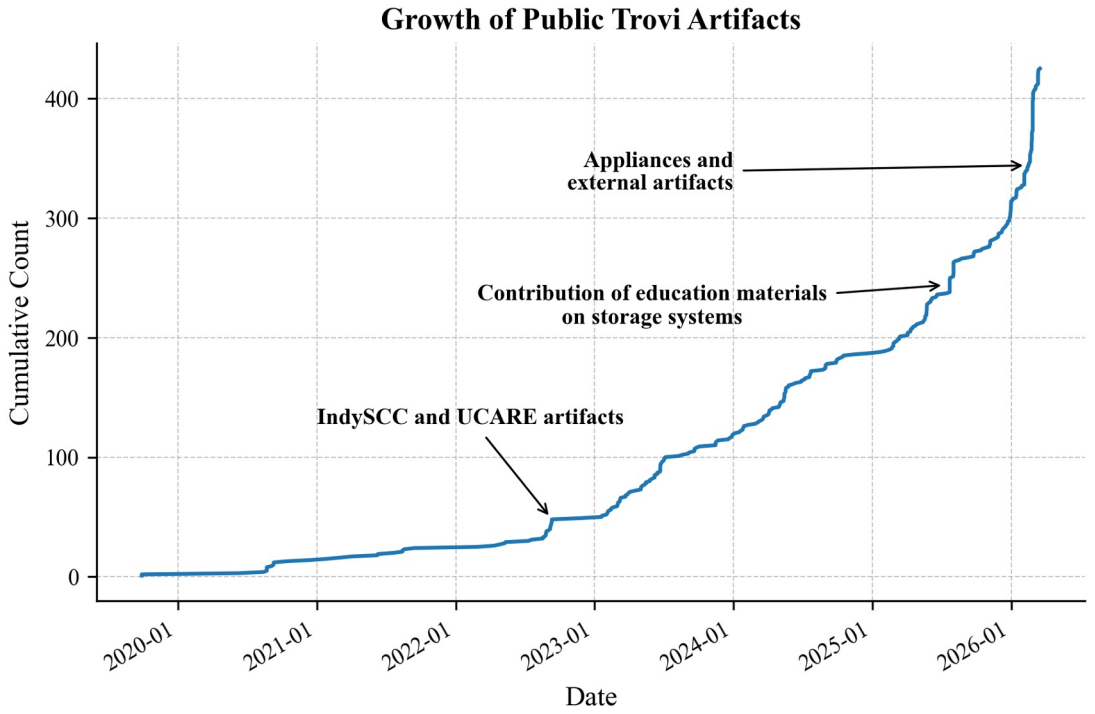
- A collection that is **close to infrastructure**
- A testbed-integrated open experiment sharing repository integrated with **multiple testbeds**
- Trovi artifacts
  - Collection of information about all the experiment
  - Connected to the testbed such that the experimental environment is easy to deploy
  - Artifacts provide metrics about usage – interesting to both authors and reviewers
- Portal to present, browse, filter, and find interesting experiments
- Open APIs: can be integrated with any testbed



Paper: "Three Pillars of Reproducibility", ReWords'23

TROVI: INFRASTRUCTURE-ENABLED ARTIFACTS

Total artifacts	425
Artifacts added in 2026	111
Artifacts added in 2025	127
Unique authors	156
Daypass enabled artifacts	27
Chameleon supported	26
Reproducible badges	42
Educational badges	55
Max access count ("launches")	3021
Mean access count	57.69
Median access count	9.00
Max unique access count	326
Mean unique access count	14.92
Median unique access count	3.00
Max unique cell execution	319

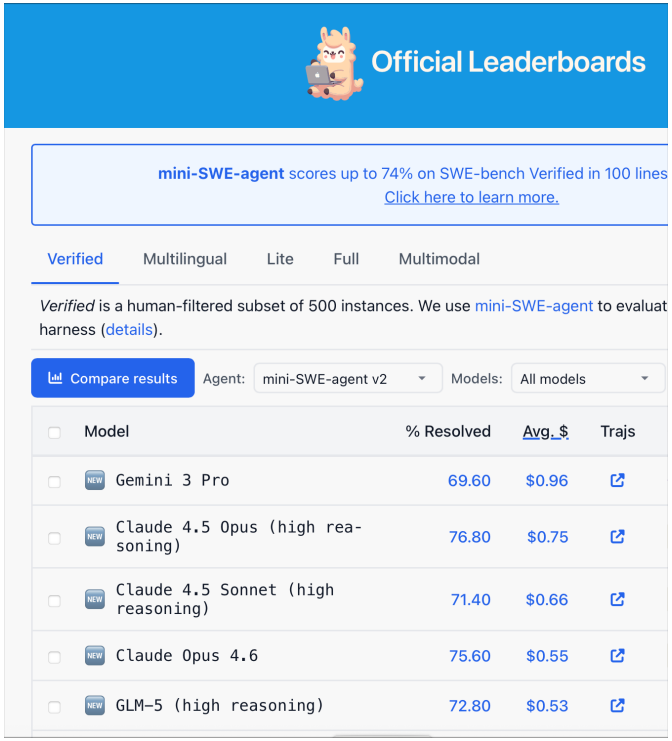


Artifacts from conferences, including: ATC, EuroSys, FAST, OSDI, SC, SOSP

Initial crawling of external artifacts found over 600 artifacts to be manually reviewed

BUT WAIT, THERE'S MORE….

- David Patterson: "For better or worse, benchmarks shape a field"
  - Better: problem characterization, focus for for innovation and research
  - Worse: Goodhart's law: "When a measure becomes a target, it ceases to be a good measure"
- SWEBench: Can LLMs resolve GitHub issues?
- SciEnvBench (Trovi collection): Can LLMs accurately re-create experimental environments?



PARTING THOUGHTS

- The shape of research infrastructure is evolving
  - More flexibility, more hardware options, more control, more interactivity
  - Slice-based resources: what used to be niche is increasingly becoming mainstream
  - Need to support cost-effective operations while also supporting evolution
- It keeps us on our toes
  - Constant evaluation is critical: does the solution still fit the problem? How is the community changing? How is usage changing? What science is produced?
  - Better understanding of how to measure the effectiveness of scientific instruments
- How do we make a better scientist?
  - By outsourcing the tedious and fostering the creative
  - Trade-off between ease of use and control just got easier – it's coming for the infrastructure near U!