Open Cirrus™ Cloud Computing Testbed

Federated Data Centers for Open Source Systems and Services Research

Roy Campbell,[5] Indranil Gupta,[5] Michael Heath,[5] Steven Y. Ko,[5]
Michael Kozuch,[3] Marcel Kunze,[4] Thomas Kwan,[6] Kevin Lai,[1] Hing Yan Lee,[2]
Martha Lyons,[1] Dejan Milojicic,[1] David O'Hallaron,[3] and Yeng Chai Soh[2]

[1]HP Labs, [2]IDA, [3]Intel Research, [4]KIT, [5]UIUC, and [6]Yahoo!

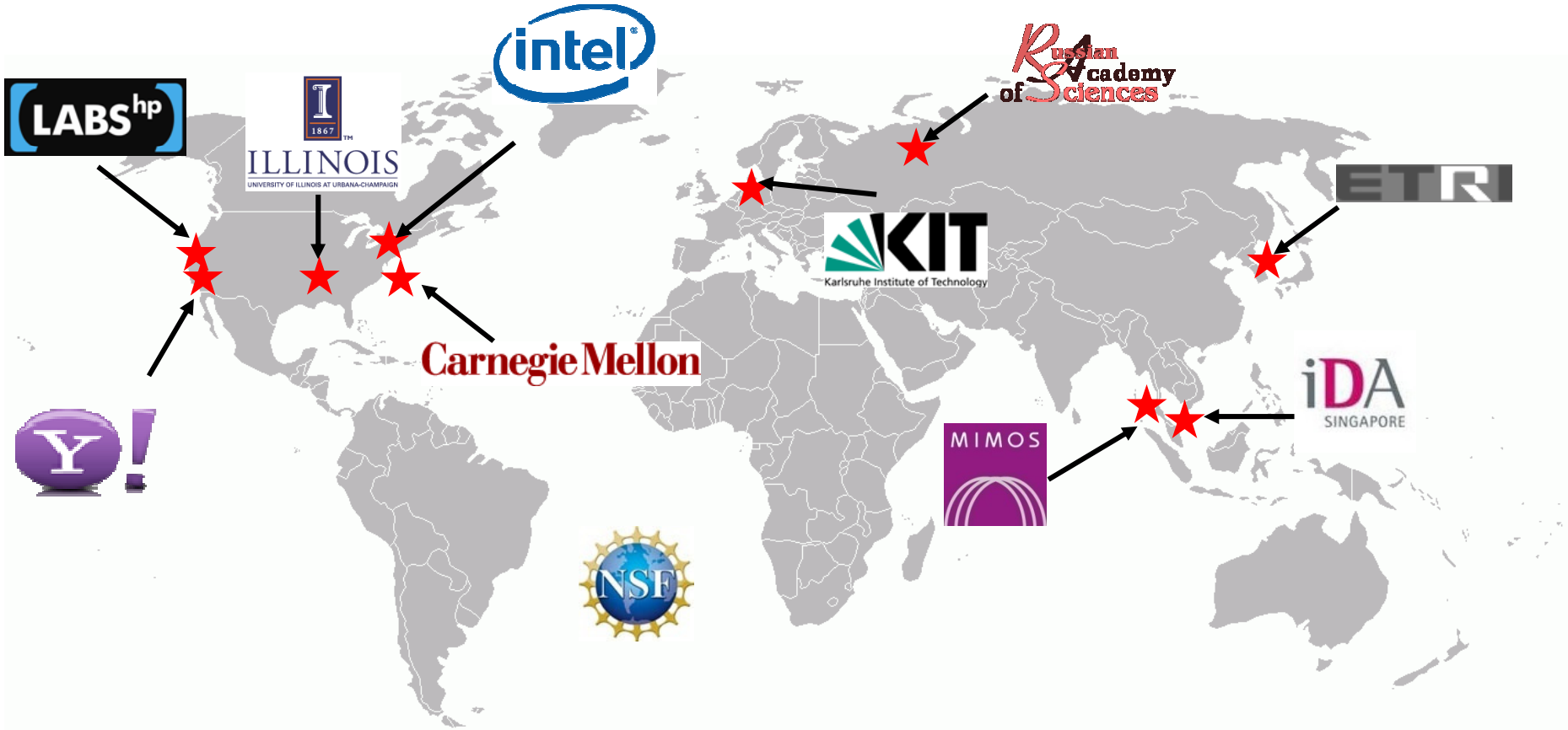
Open Cirrus™ Cloud Computing Testbed

Shared: research, **applications**, **infrastructure** (11K cores), **data sets**

Global services: sign on, monitoring, store. **Open src stack** (prs, tashi, hadoop)

Sponsored by **HP, Intel, and Yahoo!** (with additional support from **NSF**)

- 9 sites currently, target of around 20 in the next two years.



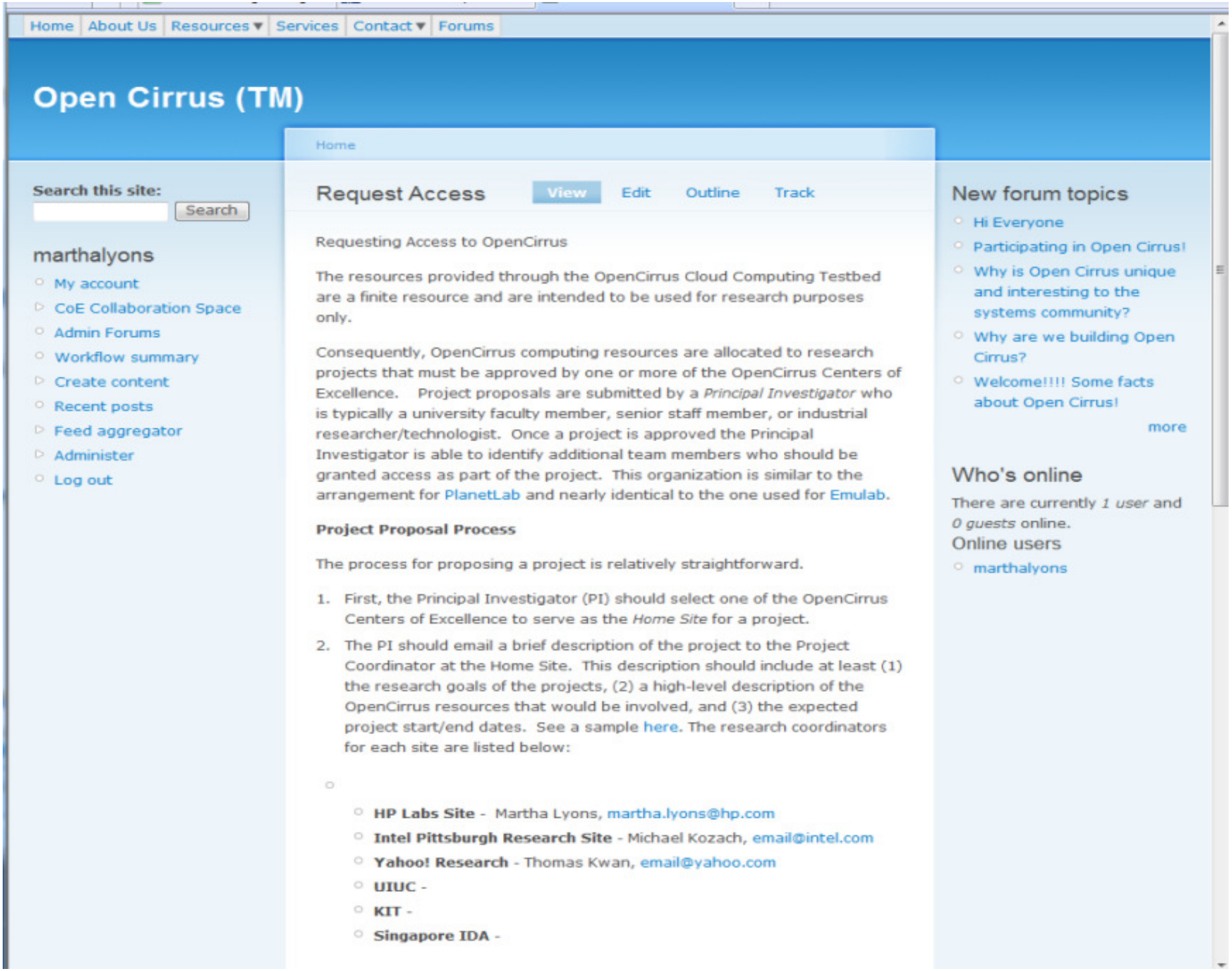
Open Cirrus

- Objectives
  - Foster systems research around cloud computing
  - Vendor-neutral open-source stacks and APIs for the cloud
  - Expose research community to enterprise level requirements
  - Provide realistic traces of cloud workloads

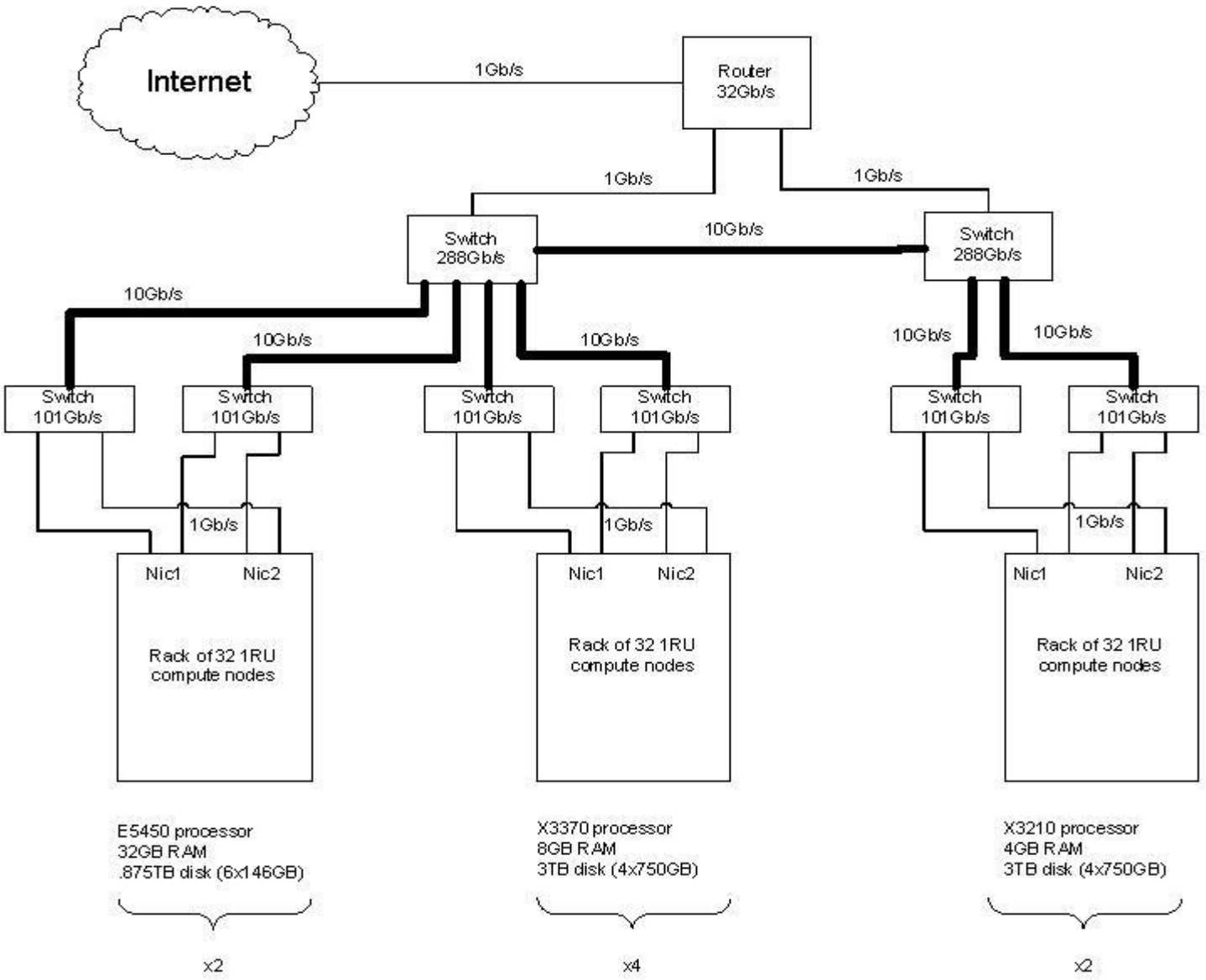
- How are we unique
  - Support for systems research and applications research
  - Federation of heterogeneous datacenters
  - Interesting data sets

Process

- Central Management Office, oversees Open Cirrus
- Governance model
  - Research team
  - Technical Team
  - New site additions
  - Support (legal (export, privacy), IT, etc.)
- Each site
  - Runs its own research and technical teams,
  - Contributes individual technologies
  - Operates some of the global services
- E.g. HP Site supports: Portal and PRS


Home
About Us
Resources
Services
Contact
Forums
Open Cirrus (TM)
Home
Search this site:
Search
marthalyons
My account
CoE Collaboration Space
Admin Forums
Workflow summary
Create content
Recent posts
Feed aggregator
Administer
Log out
Request Access
View
Edit
Outline
Track
Requesting Access to OpenCirrus
The resources provided through the OpenCirrus Cloud Computing Testbed are a finite resource and are intended to be used for research purposes only.
Consequently, OpenCirrus computing resources are allocated to research projects that must be approved by one or more of the OpenCirrus Centers of Excellence. Project proposals are submitted by a Principal Investigator who is typically a university faculty member, senior staff member, or industrial researcher/technologist. Once a project is approved the Principal Investigator is able to identify additional team members who should be granted access as part of the project. This organization is similar to the arrangement for PlanetLab and nearly identical to the one used for Emulab.
Project Proposal Process
The process for proposing a project is relatively straightforward.
1. First, the Principal Investigator (PI) should select one of the OpenCirrus Centers of Excellence to serve as the Home Site for a project.
2. The PI should email a brief description of the project to the Project Coordinator at the Home Site. This description should include at least (1) the research goals of the projects, (2) a high-level description of the OpenCirrus resources that would be involved, and (3) the expected project start/end dates. See a sample here. The research coordinators for each site are listed below:
HP Labs Site - Martha Lyons, martha.lyons@hp.com
Intel Pittsburgh Research Site - Michael Kozach, email@intel.com
Yahoo! Research - Thomas Kwan, email@yahoo.com
UIUC -
KIT -
Singapore IDA -
New forum topics
Hi Everyone
Participating in Open Cirrus!
Why is Open Cirrus unique and interesting to the systems community?
Why are we building Open Cirrus?
Welcome!!!! Some facts about Open Cirrus!
more
Who's online
There are currently 1 user and 0 guests online.
Online users
marthalyons

HP Labs Cirrus Cluster topology



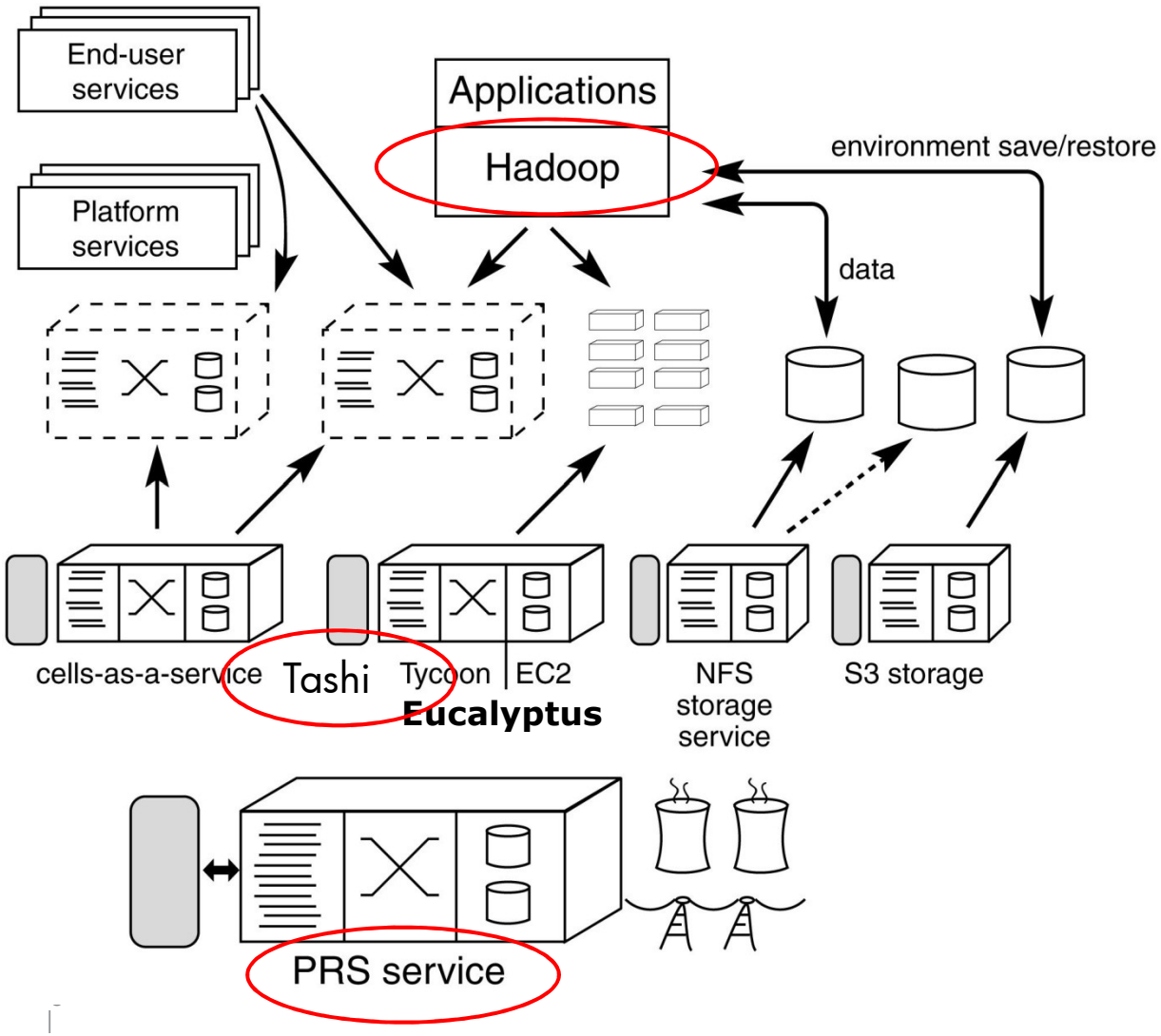
				Totals
Nodes/cores:	64/256	128/512	64/256	256/1024
RAM (GB):	2048	1024	256	3328
Storage (TB):	56	384	192	632
Spindles:	384	512	256	1152

Open Cirrus Sites

Site	Characteristics							
	#Cores	#Servers	Public partition	Memory Size	Storage Size	Spindles	Network	Focus
HP	1,024	256	178	3.3TB	632TB	1152	10G internal 1Gb/s x-rack	Hadoop, Cells, PRS, scheduling
IDA	2,400	300	100	4.8TB	43TB+ 16TB SAN	600	1Gb/s	Apps based on Hadoop, Pig
Intel	1060	155	145	1.16TB	353TB local 60TB attach	550	1Gb/s	Tashi, PRS, MPI, Hadoop
KIT	2048	256	128	10TB	1PB	192	1Gb/s	Apps with high throughput
UIUC	1024	128	64	2TB	~500TB	288	1Gb/s	Datasets, cloud infrastructure
Yahoo	3200	480	400	2.4TB	1.2PB	1600	1Gb/s	Hadoop on demand

LABS hp

Open Cirrus Software Stack



Cloud application services

Virtual Resource Sets

Cloud infrastructure services

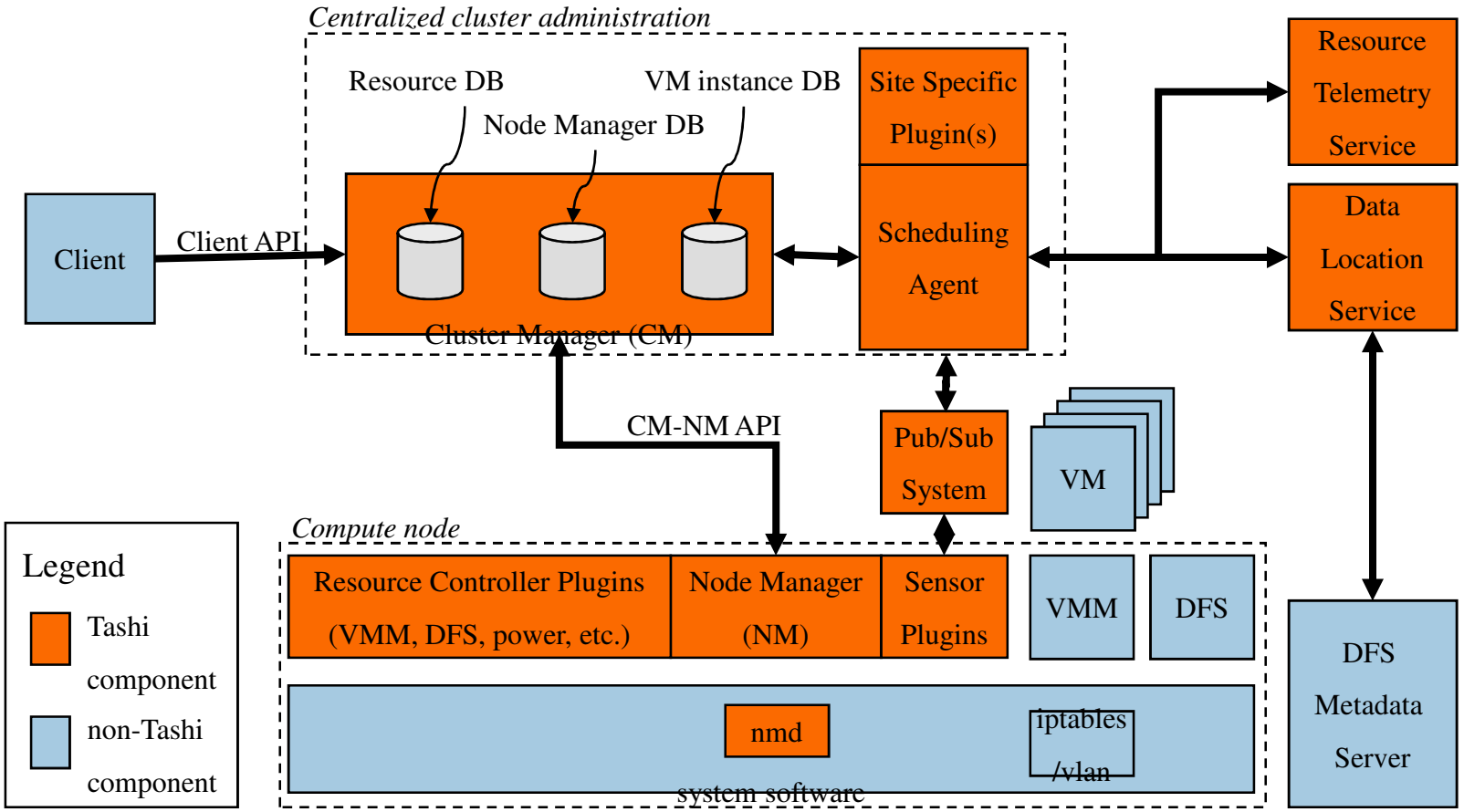
IT infrastructure layer

(Physical Resource Sets)

Physical Resource Sets (PRS)

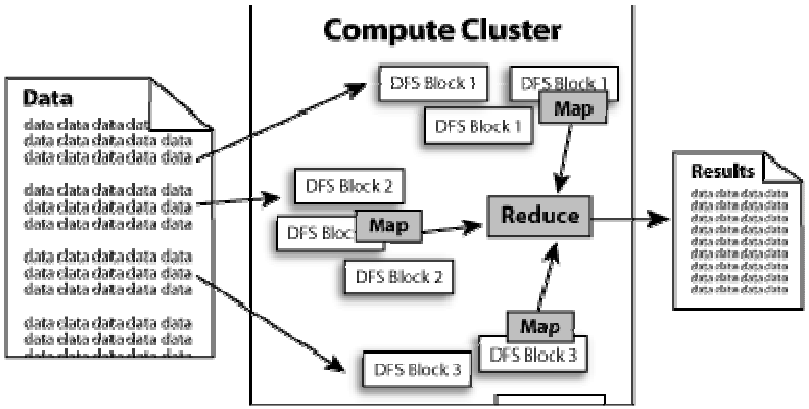
- PRS service goals
  - Provide mini-datacenters to researchers
  - Isolate experiments from each other
  - Stable base for other research
- PRS service approach
  - Allocate sets of physical co-located nodes, isolated inside VLANs
  - Start simple, add features as we go
  - Base to implement virtual resource sets
- Hardware as a Service (HaaS)

Tashi Software Architecture



Programming the Cloud: Hadoop

- An open-source Apache software foundation project sponsored by Yahoo!
  - http://wiki.apache.org/hadoop/ProjectDescription
  - reproduce the proprietary software infrastructure developed by Google
- Provides a parallel programming model (MapReduce), a distributed file system, and a parallel database
  - http://en.wikipedia.org/wiki/Hadoop
  - http://code.google.com/edu/parallel/mapreduce-tutorial.html



How do users get access to Open Cirrus sites?

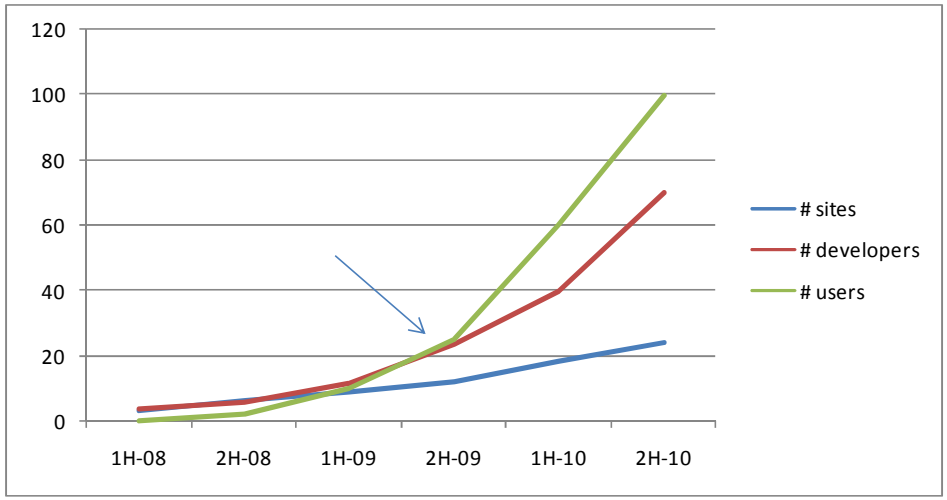
- Project PIs apply to each site separately
- Contact email addresses on the Open Cirrus portal
  - http://opencirrus.org
- Each Open Cirrus site decides which users and projects get access to its site
- A *global sign on* for all sites
  - Users are able to login to each OpenCirrus site for which they are authorized using the same login and password.

What kinds of research projects are Open Cirrus sites looking for?

- Open Cirrus™ is seeking research in the following areas (different centers will weight these differently)
  - Datacenter federation
  - Datacenter management
  - Web services
  - Data-intensive applications and systems
  - Hadoop map-reduce applications

- The following kinds of projects are of less interest
  - Traditional HPC application development
  - Production applications that just need lots of cycles
  - Closed source system development

Metrics of Success



- Community
  - Technology used
  - # Sites, Projects, (Vibrant) Users
  - Research Productivity (Shared Cost of Research), # papers published
  - Cross-collaboration (Portal traffic)
  - # New open source components
- Technical
  - Utilization of Open Cirrus, TCO
  - Ease of use (e.g. provision 50% of OC nodes in < 30sec)
  - Federation transparency/adoption
  - Reliability

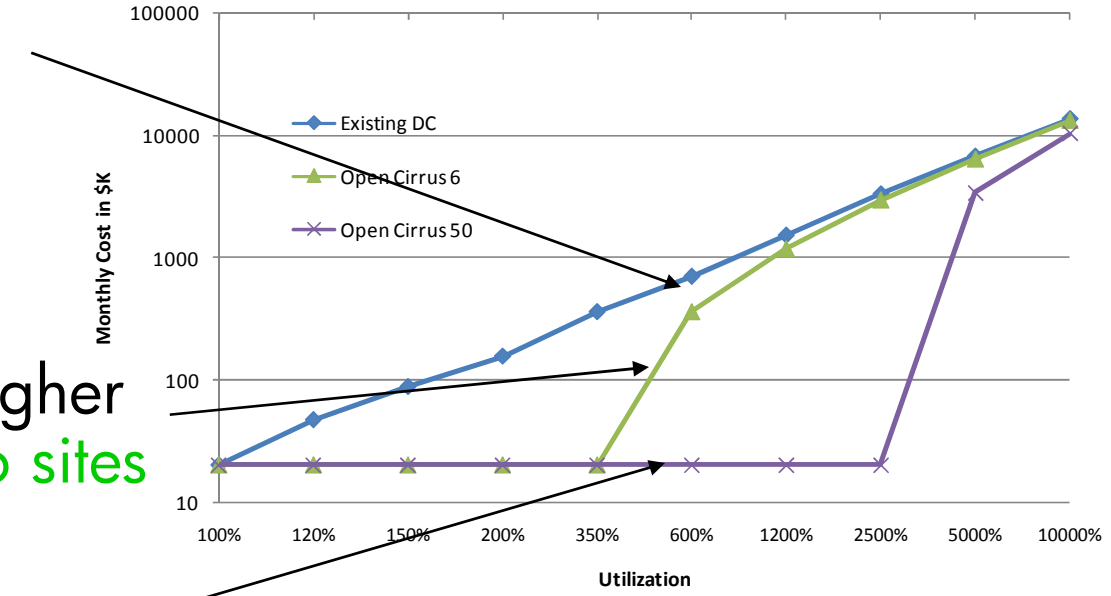
LABS hp

Single site Cloud: to Outsource or Own?

- Medium-sized organization: wishes to run a service for *M* months
  - Service requires 128 servers (1024 cores) and 524 TB
  - Same as UIUC cloud site
- Outsource (e.g., via AWS): monthly cost
  - Storage ~ $62 K
  - Total ~ $136 K (using 0.45:0.0.4:0.15 split for hardware:power:network)
- Own: monthly cost
  - Storage ~ $349 K / *M*
  - $ 1555 K / *M* + 7.5 K (includes 1 sysadmin / 100 nodes)
- Breakeven analysis: **more preferable to own if**:
  - *M* > 5.55 months (storage)
    - Not surprising: Cloud providers benefit monetarily most from storage
  - *M* > 12 months (overall)
- With underutilization of x%, still more preferable to own if:
  - x > 33.3%
  - Even with CPU util of 20%, storage > 47% makes owning preferable

Federation Economics

- Federation can help contain demand overflow within itself
- Cost of outsourcing overflow to AWS is higher than to federation of 6 sites
- Cost reduces with size of federation increasing to 50



LABS hp

Open Cirrus v. Other Testbeds

	Testbeds							
	Open Cirrus	**IBM/Google**	**TeraGrid**	**PlanetLab**	**EmuLab**	**Open Cloud Consortium**	**Amazon EC2**	**LANL/NSF cluster**
Type of research	Systems & applications	Data-intensive applications research	Scientific applications	Systems and services	Systems	interoperab. across clouds using open APIs	Commer. use	Systems
Approach	Federation of heterog. data centers	A cluster supported by Google and IBM	Multi-site hetero clusters super comp.	A few 100 nodes hosted by research instit.	A single-site cluster with flexible control	Multi-site heterogeneous clusters	Raw access to virtual machines	Re-use of LANL's retiring clusters
Participants	HP, Intel, IDA, KIT, UIUC, Yahoo!	IBM, Google, Stanford, U.Washington, MIT	Many univ. & orgs	Many univ & organizations	University of Utah	4 centers –	Amazon	CMU, LANL, NSF
Distribution	6 sites	1 site	11 partners in US	> 700 nodes world-wide	>300 nodes univ@Utah	480 cores, distributed in four locations		1000s of older, still useful nodes at 1 site

Open Cirrus Research Summary

HP	Intel	
• Mercado • Policy Aware Data Mgmt • Wikipedia Mining & tagging • SPARQL Query over Hadoop (UTD) • N-tier App Benchmark (GaTech)	• Everyday Sensing and Perception • SLIPstream/Sprout • Parallel Machine Learning • NeuroSys • Computational Health • FastBeat (w/France Telecom)	*Cloud application frameworks and services*
• Economic Cloud Stack • Parallel Data Series	• Tashi (with CMU, Yahoo)	*Cloud infrastructure services*
• OpenNet • Exascale Data Center	• PRS (with HP)	*IT infrastructure layer*

Summary

- Cloud is creating a new paradigm in computing
  - Flexible and elastic resource provisioning
  - Economy of scale makes it attractive
  - Move from manufacture towards industrialization of IT (Everything as a Service)
- OpenCirrus offers interesting R&D opportunities
  - Cloud systems and applications research and development
  - Interesting data sets and federation of heterogeneous data centers
- OpenCirrus workshop at HP Palo Alto on June 8/9 has links to a lot of materials

LABS hp