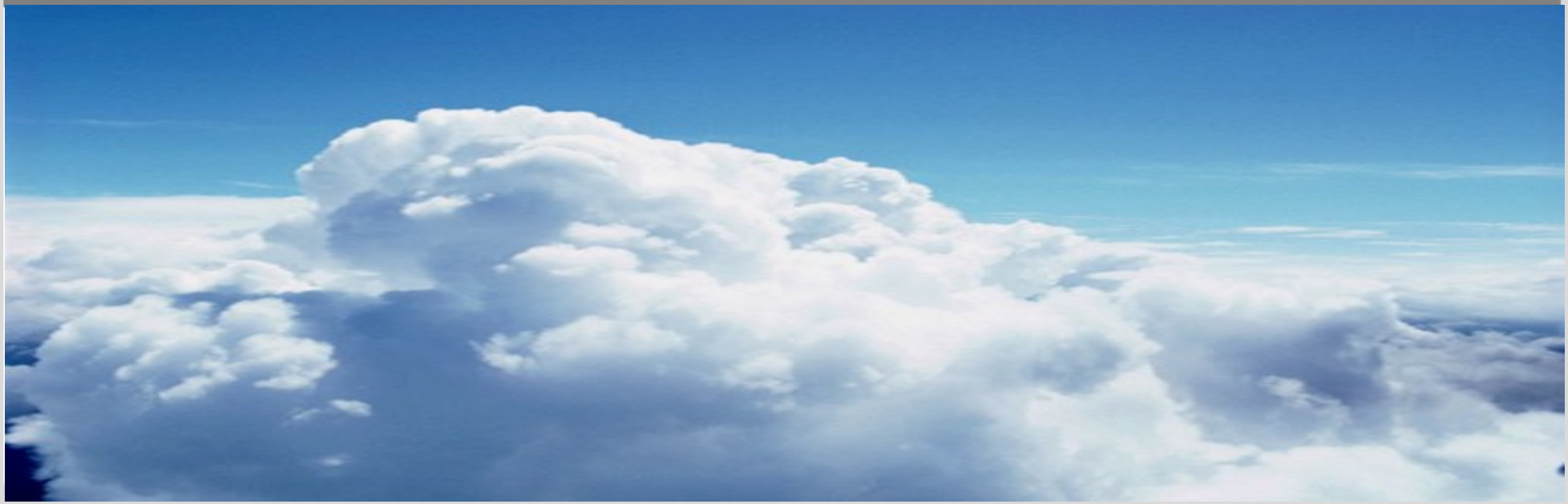
The Open Cirrus Project

Towards an Open-source Cloud Stack

Marcel Kunze, Steinbuch Centre for Computing



Steinbuch Center for Computing (SCC)

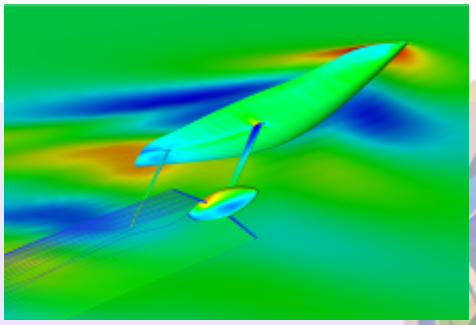
Karlsruhe Institute of Technology

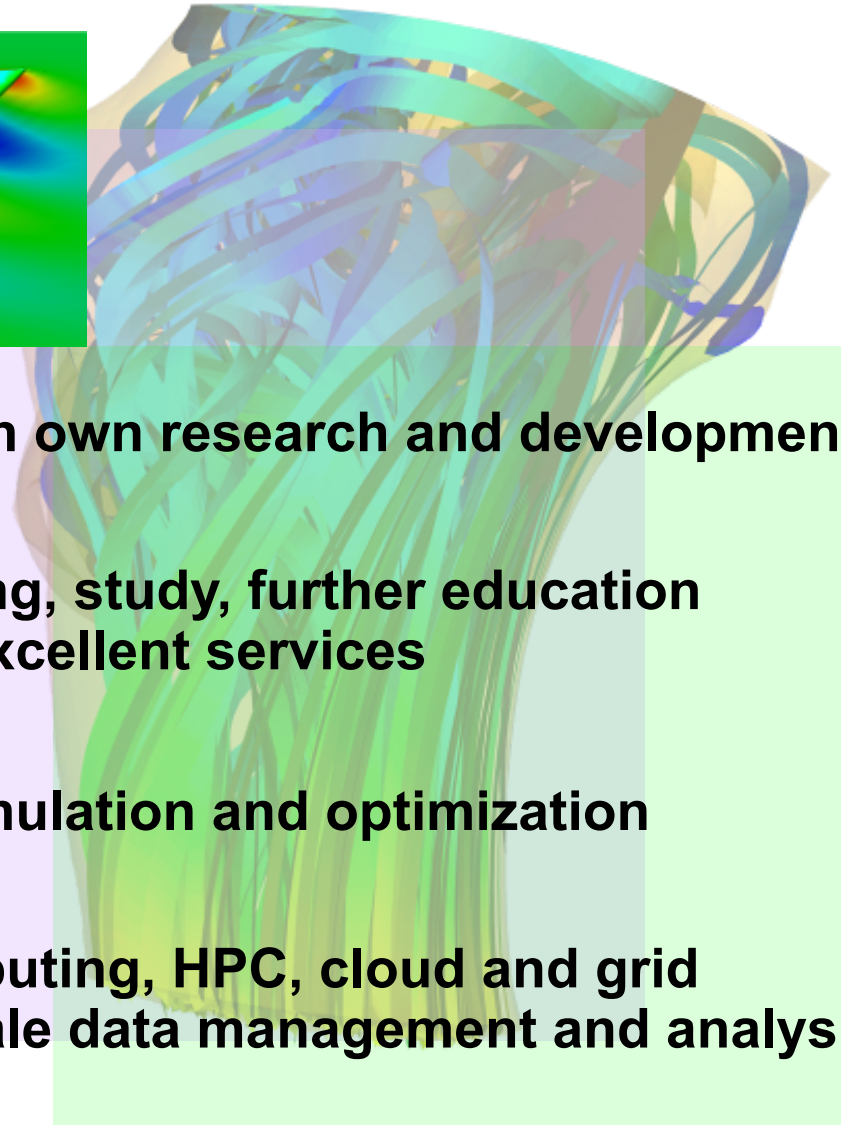
SCC at
Karlsruhe University

SCC at Research Center

- **Founded on January 1st, 2008**
- **Information technology center of KIT**
- **Merger of the computing center of Karlsruhe University and IWR at FZK**
- **One of the largest scientific computing centers in Europe**

Mission of the SCC





- IT-Services under one roof with own research and development
- Promotion of research, teaching, study, further education and administration at KIT by excellent services
- Major center for modelling, simulation and optimization
- Leading role in scientific computing, HPC, cloud and grid computing as well as large scale data management and analysis

The Cloud is just a new Computing Paradigm

- 70's: Mainframe
- 80's: PC
- 90's: Workstation
- 00's: Grid
- 10's: Cloud

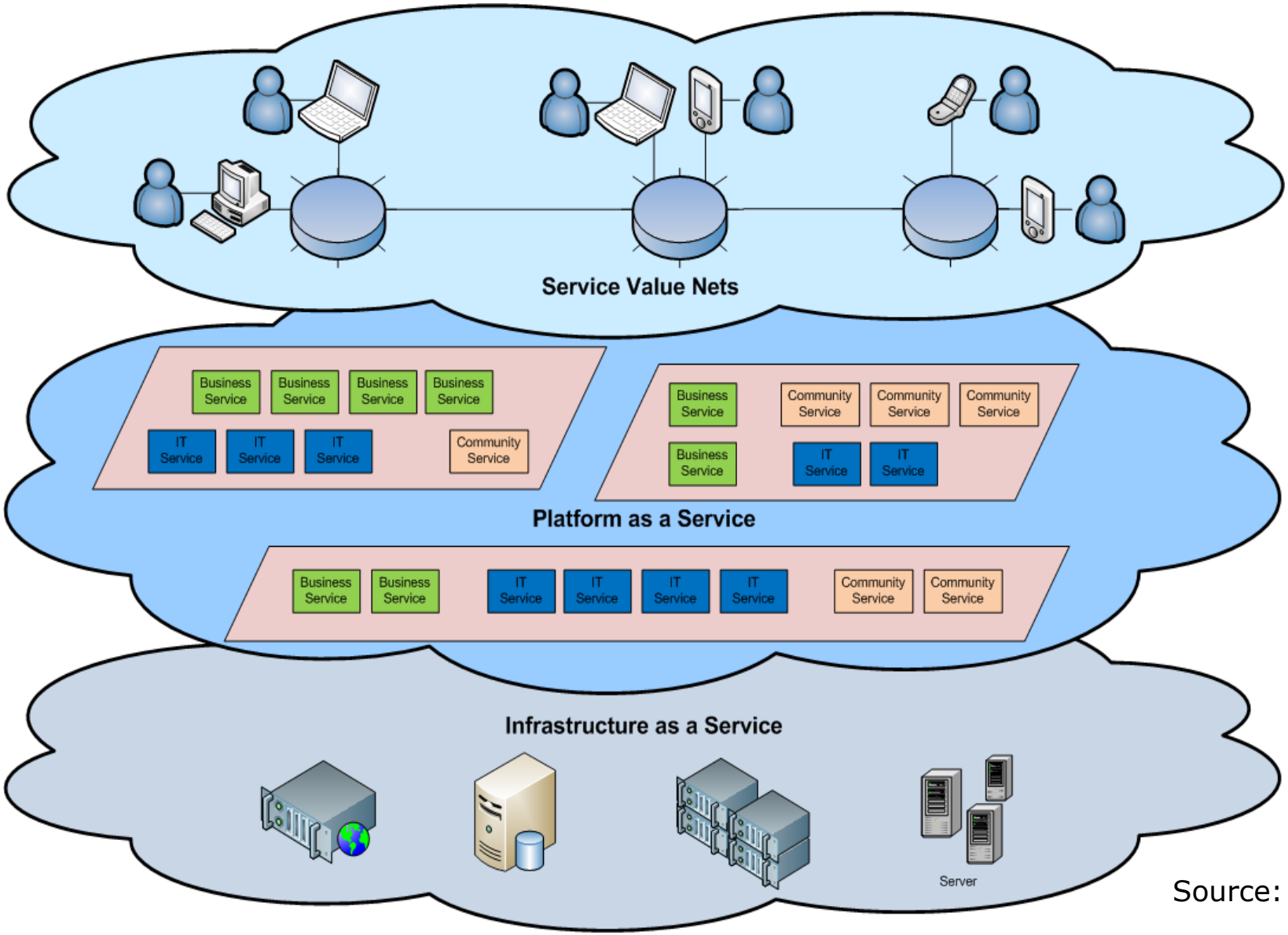
From http://blogs.zdnet.com/Hinchcliffe

Cloud Computing: A possible Definition

"Building on compute and storage virtualization, and leveraging the modern Web, **Cloud Computing** provides scalable, network-centric, abstracted IT infrastructure, platforms, and applications as on-demand services that are billed by consumption."

C.Baun, M.Kunze, J.Nimis, S.Tai: Cloud Computing, Springer 2009

Cloud Architecture: Everything as a Service

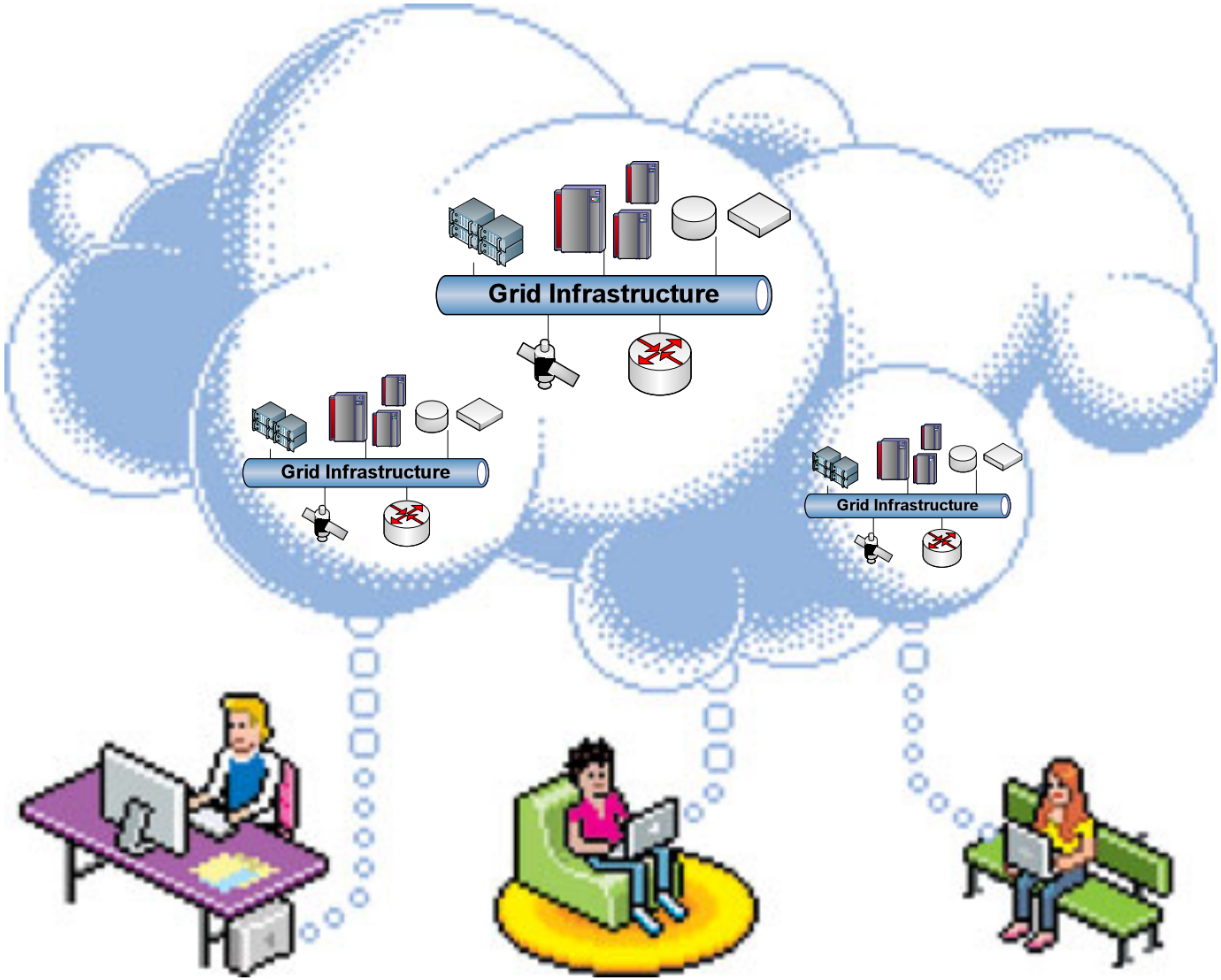


Source: S.Tai

Cloud Myths

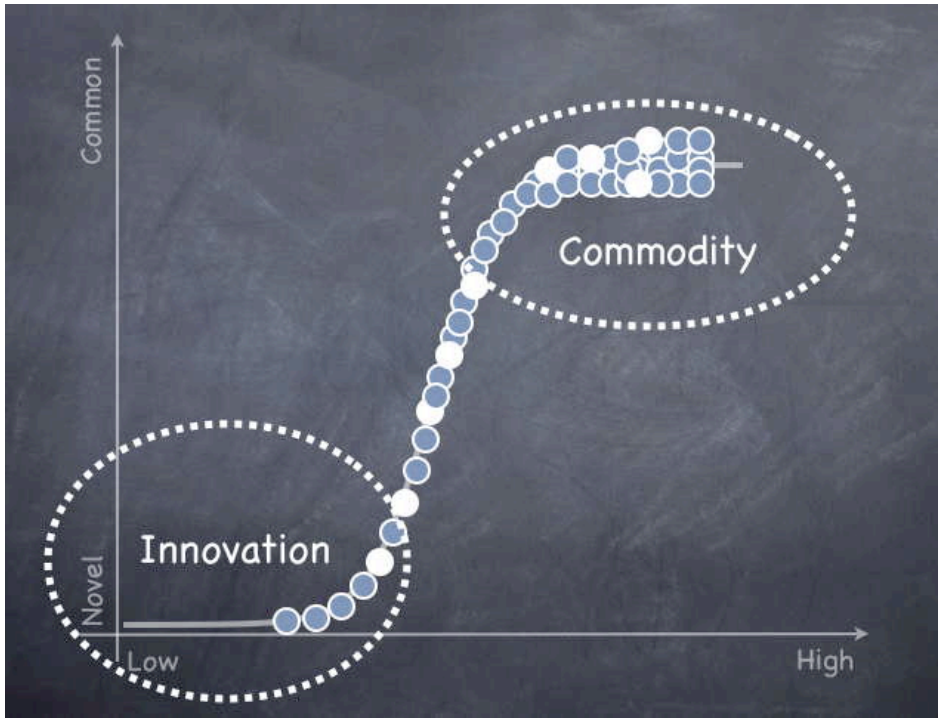
- **Cloud computing infrastructure is just a web service interface to operating system virtualization.**
  - "I'm running Xen in my data center – I'm running a private cloud."
  - Cloud needs more: Automation, SLAs, business model,…
- **Cloud computing imposes a significant performance penalty over "bare metal" provisioning.**
  - "I won't be able to run a private cloud because my users will not tolerate the performance hit."
  - With modern hardware, virtualized services are sometimes even faster
- **Clouds and Grids are equivalent.**
  - "In the mid 1990s, the term grid was coined to describe technologies that would allow consumers to obtain computing power on demand."
  - Grids are not self-service, do not grant privileges to users, have a de-centralized management and often are missing business models

Grid as a Service



Open-source Cloud Infrastructure

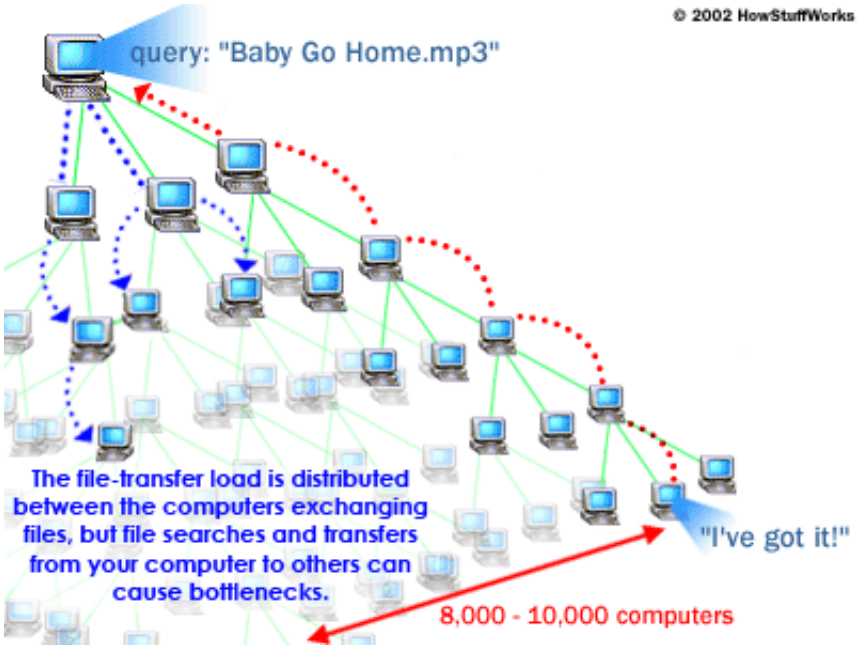
- Cloud computing is primarily a commercial endeavor
- What are the options for open-source to stay competitive with cloud computing?
- Open-source is a driver to move technologies from innovation to commodity and towards utility services



First Option: Cooperation

- Create open-source infrastructure for clouds
- Return innovation and best practices to the community
- Apply open business models to all components
- Everybody can participate as resource provider and consumer
- Possibly evolving into the largest single cloud ecosystem in the world

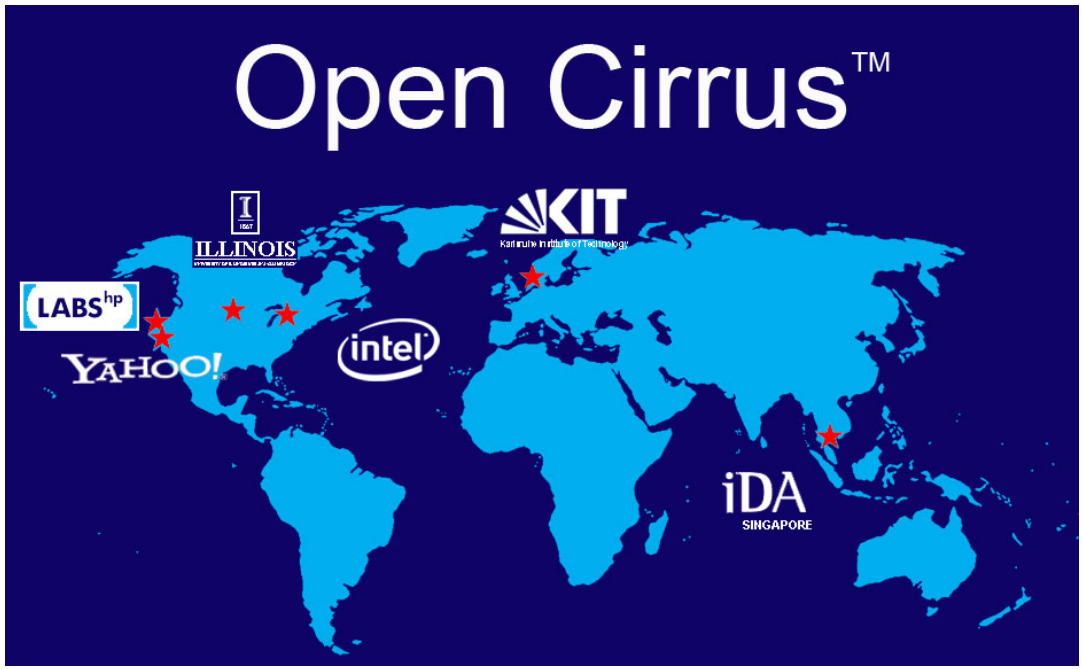
- BitTorrent as an example



Second Option: Federation

- Take advantage of standards
- Interoperability between similar data centers
- Enterprise virtual private clouds
- Open-source drives standardization and interoperability
- Overcome vendor lock-in and create a market

- Open Cirrus as an example



Open Cirrus Cloud Computing Research Testbed

http://opencirrus.org

- **An open, internet-scale global testbed for cloud computing research**
  - A tool for collaborative research
  - Focus: data center management & cloud services
- **Resources**
  - Multi-continent, multi-datacenter, cloud computing system
  - Federated “Centers of Excellence” around the globe
    - each with 100–400+ nodes and up to ~2PB storage
    - and running a suite of cloud services
- **Structure**
  - Sponsors: HP Labs, Intel Research, Yahoo!
  - Partners: UIUC, Singapore IDA, KIT, NSF
  - New partners: ETRI, MIMOS, RAS
  - Members: System and application development
- **Great opportunity for cloud systems research**
  - Accepts research proposals
  - Apply through website

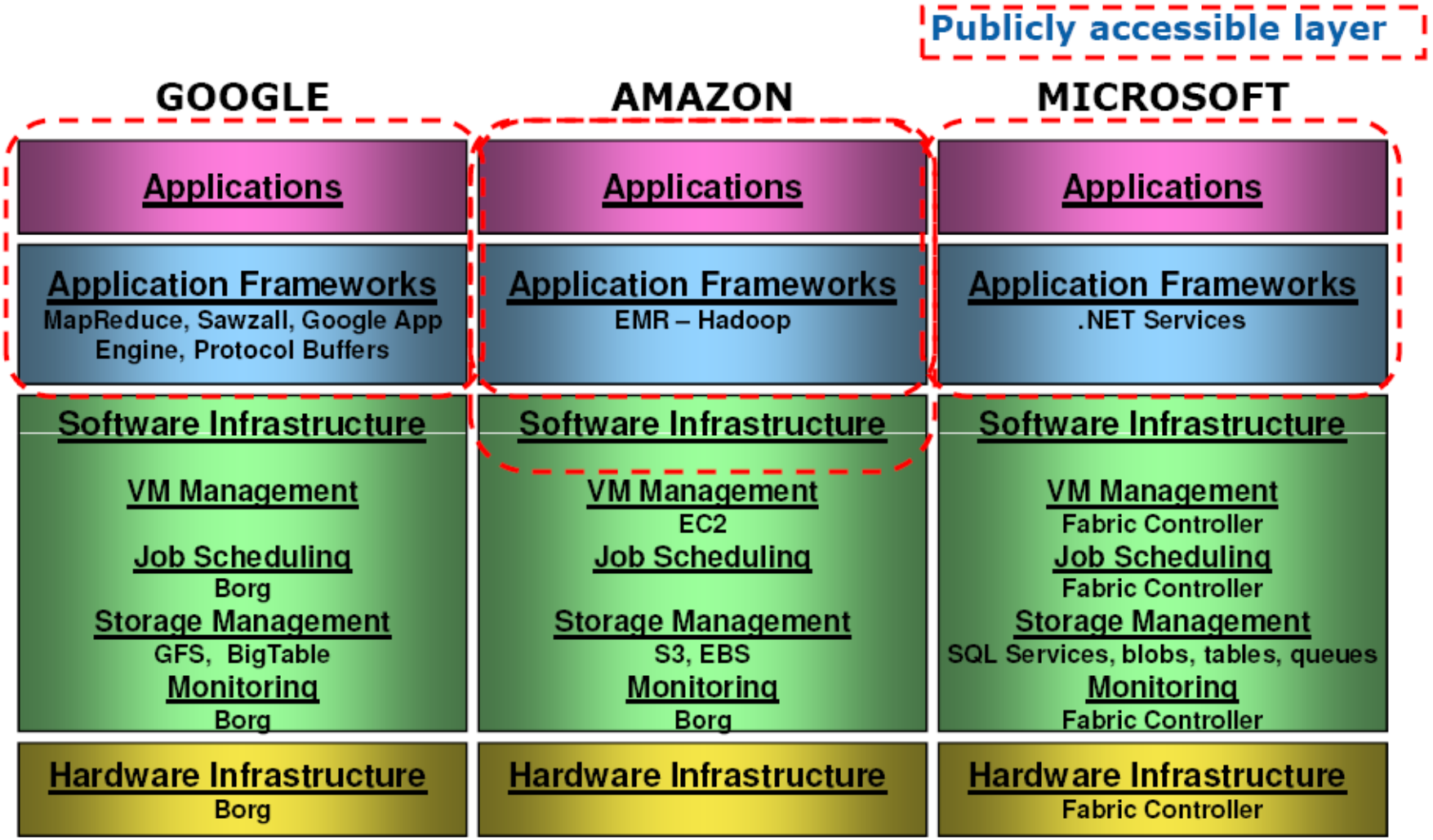
Cloud Systems Research

- **Open Cirrus is seeking research in the following areas:**
  - Datacenter federation
  - Datacenter management
  - Web services
  - Data-intensive applications and systems

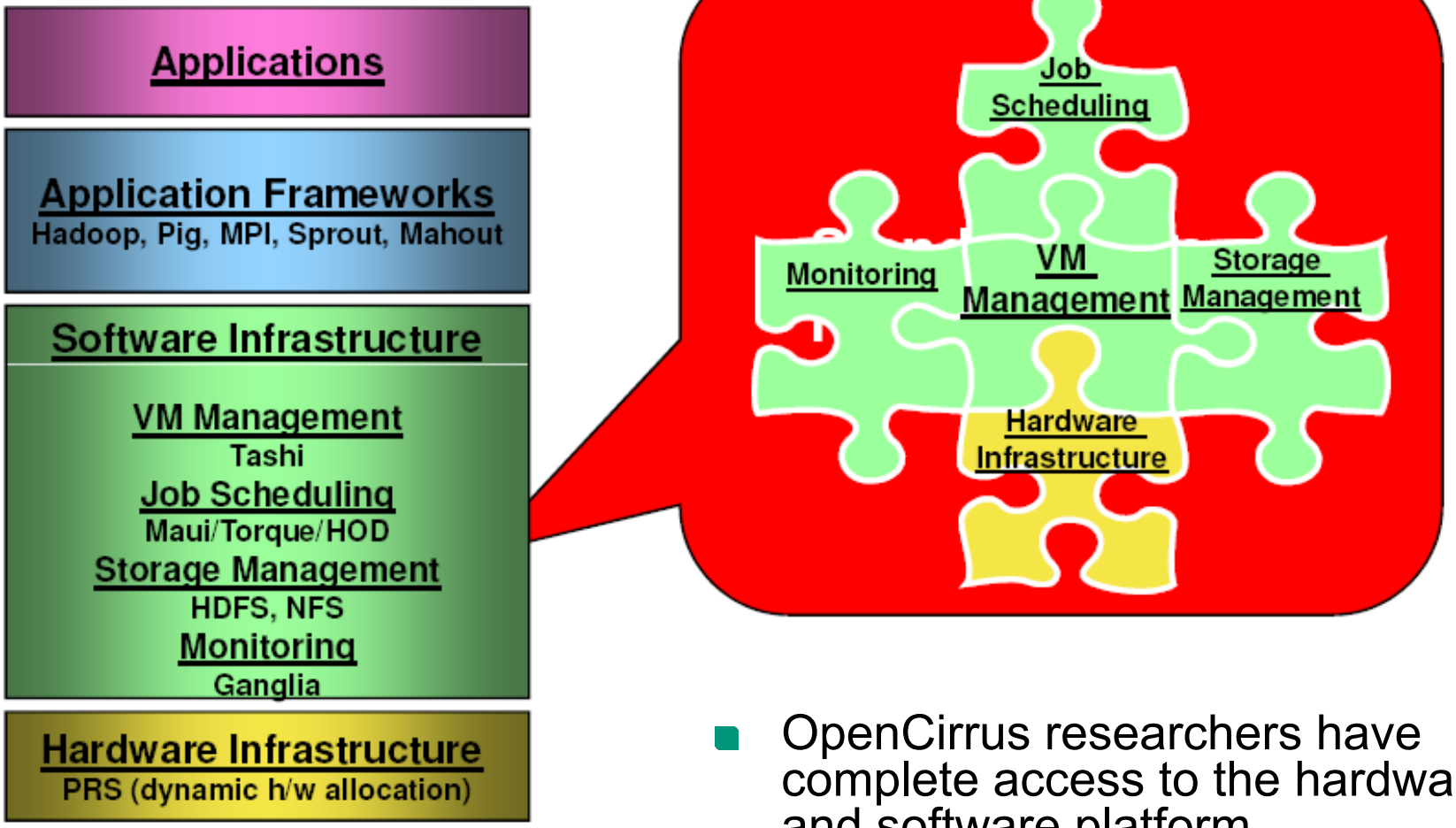
- **Research requirements**
  - Perform experiments also on a low system level
  - Flexible cloud computing framework
  - Compare different methodologies and implementations

- **Simple, transparent, controllable cloud computing infrastructure**

Proprietary Cloud Computing Stacks

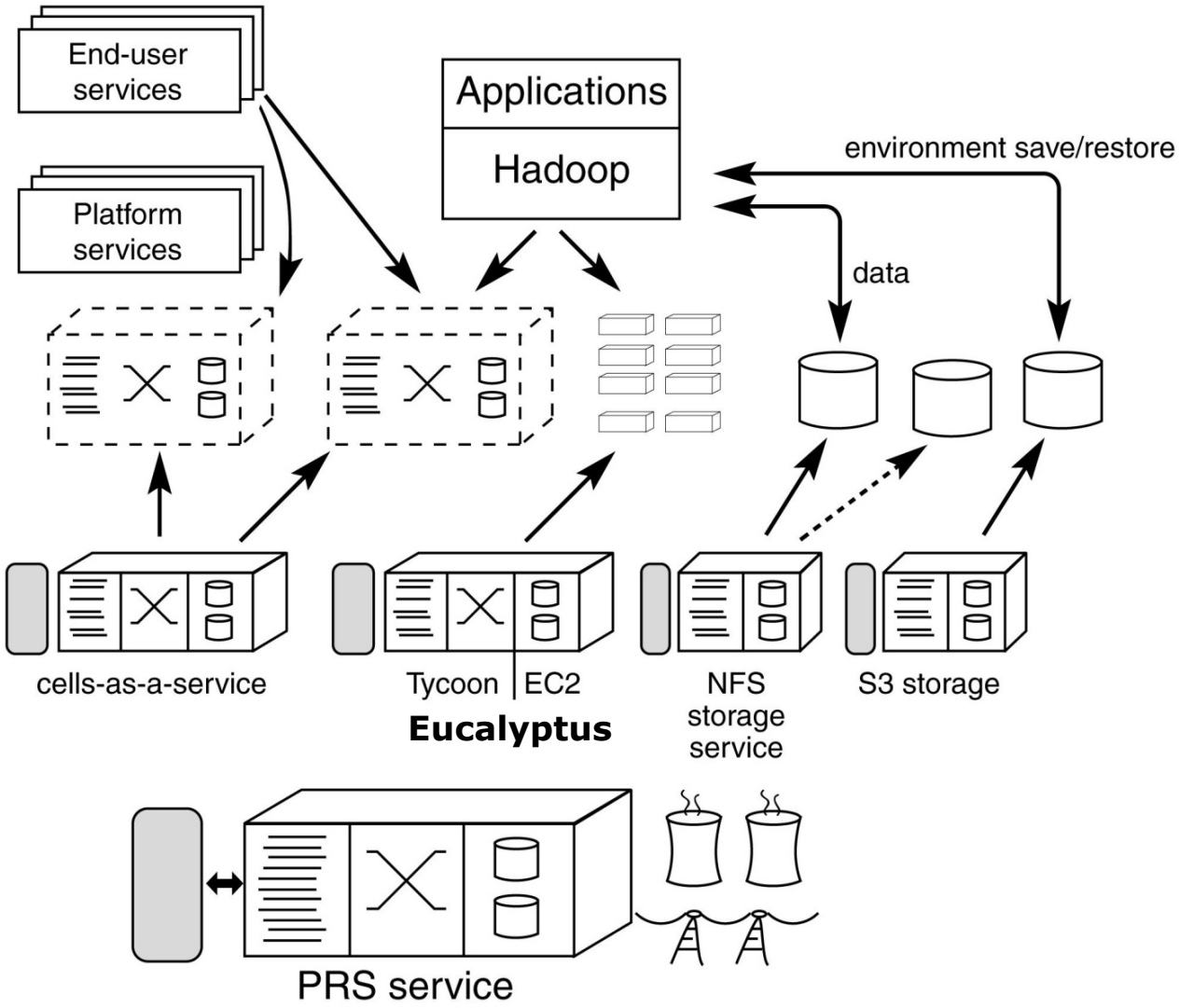


Open-source Cloud Stack



- OpenCirrus researchers have complete access to the hardware and software platform

Open Cirrus Blueprint



Cloud Application Services

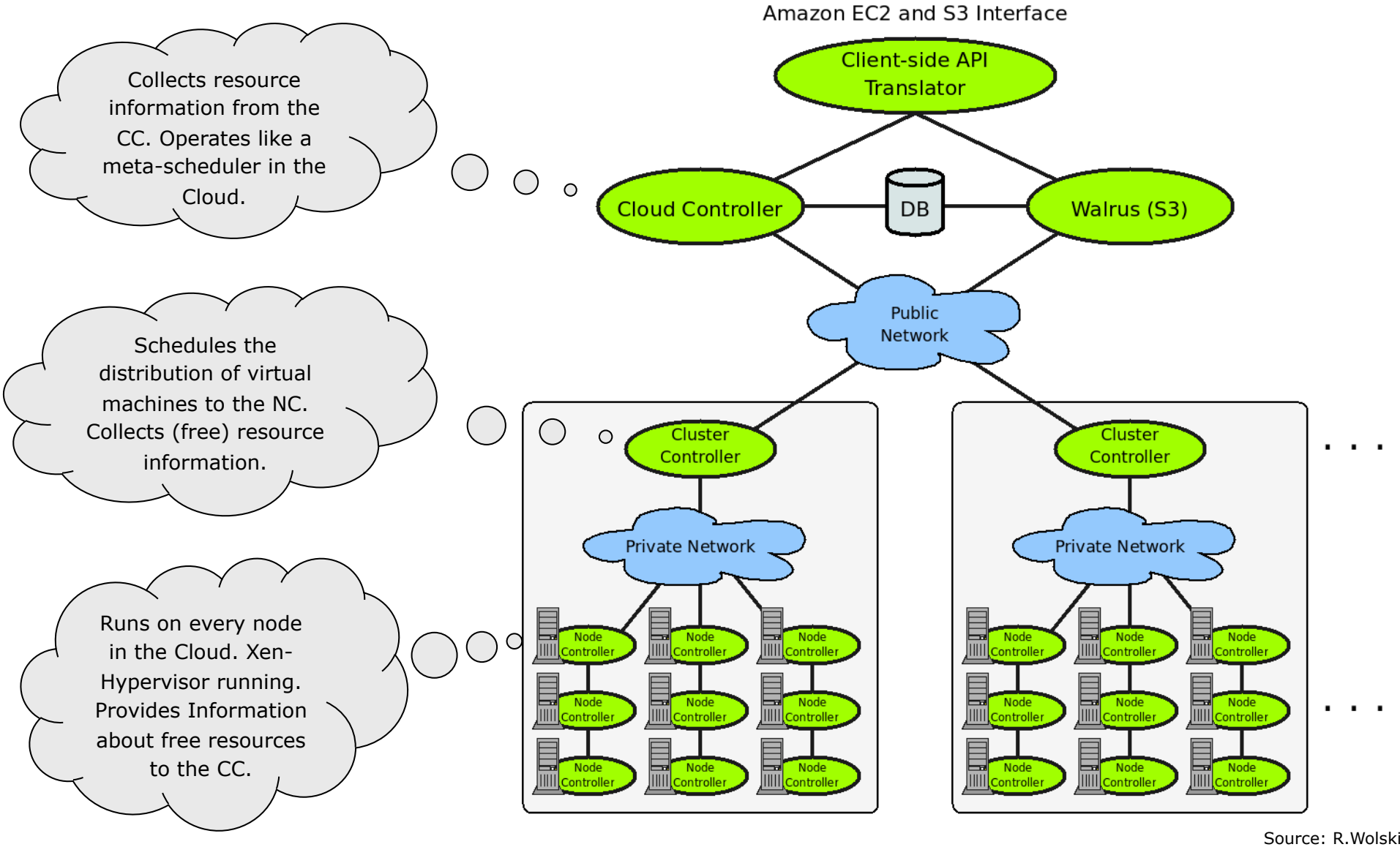
Virtual Resource Sets

Cloud Infrastructure Services

IT-Infrastructure Layer (Physical Resource Sets)

Eucalyptus: A potential VRS

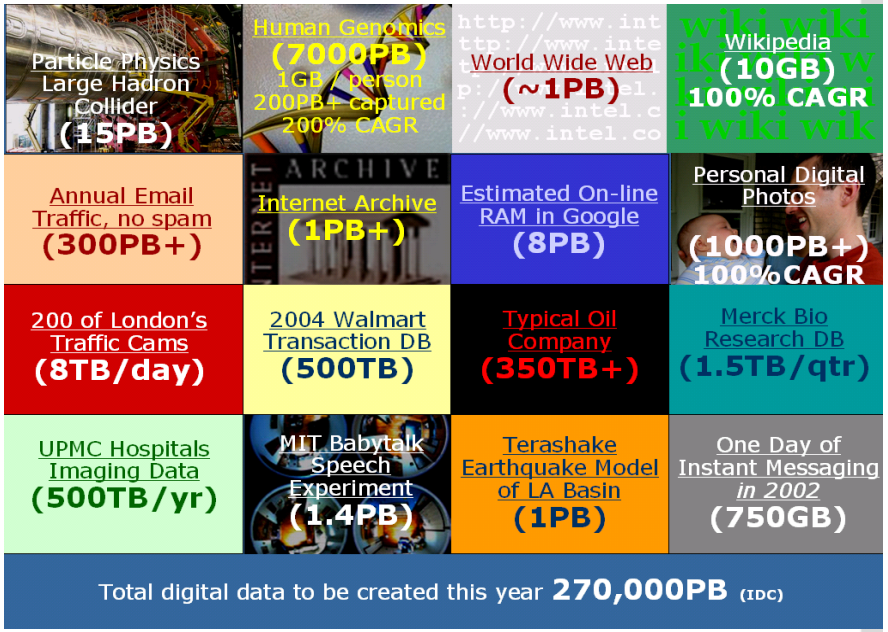
http://eucalyptus.cs.ucsb.edu



Source: R.Wolski

Big Data

- Interesting applications are *data hungry*
- The data grows over time
- The data is immobile
  - 100 TB @ 1Gbps ~= 10 days
- Compute comes to the data
- Big Data clusters are the new libraries

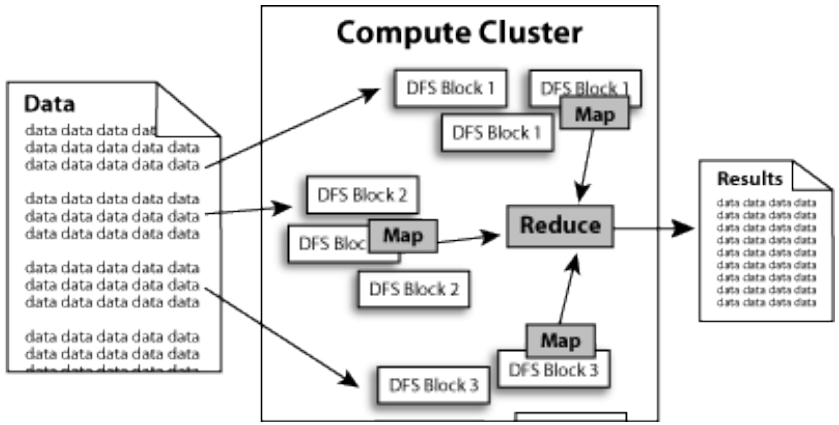


(J. Campbell, et al., Intel Research Pittsburgh, 2007)

The value of a cluster is its data

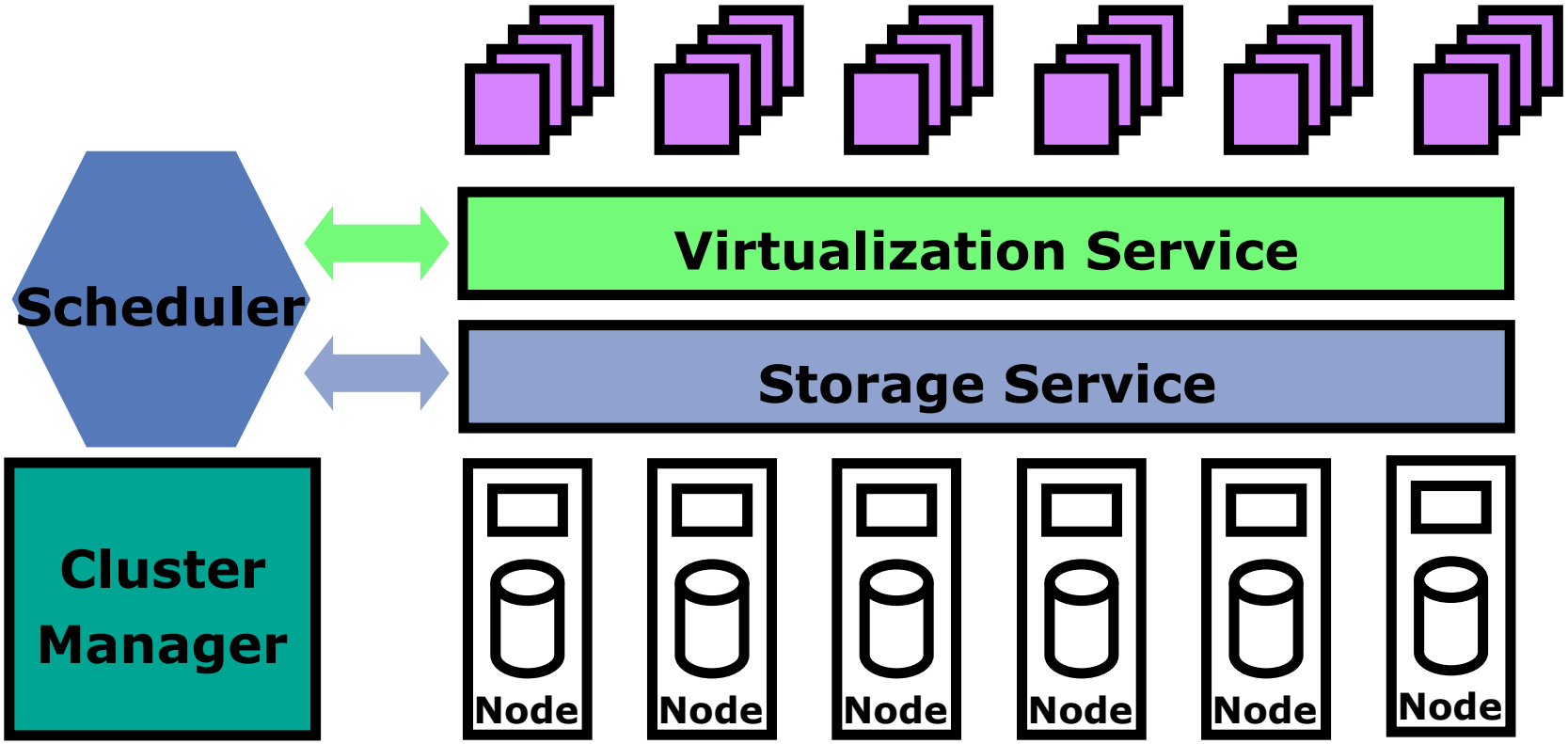
Programming the Cloud: Hadoop

- **Reproduce the proprietary software infrastructure developed by Google as open-source**
  - An Apache software foundation project sponsored by Yahoo!
- **Provides a parallel programming model (MapReduce), a distributed file system, and a parallel database**
  - Especially well suited for analysis of “Big Data”
  - The largest Hadoop cluster at Yahoo! comprises 32000 cores and 16 Petabytes storage



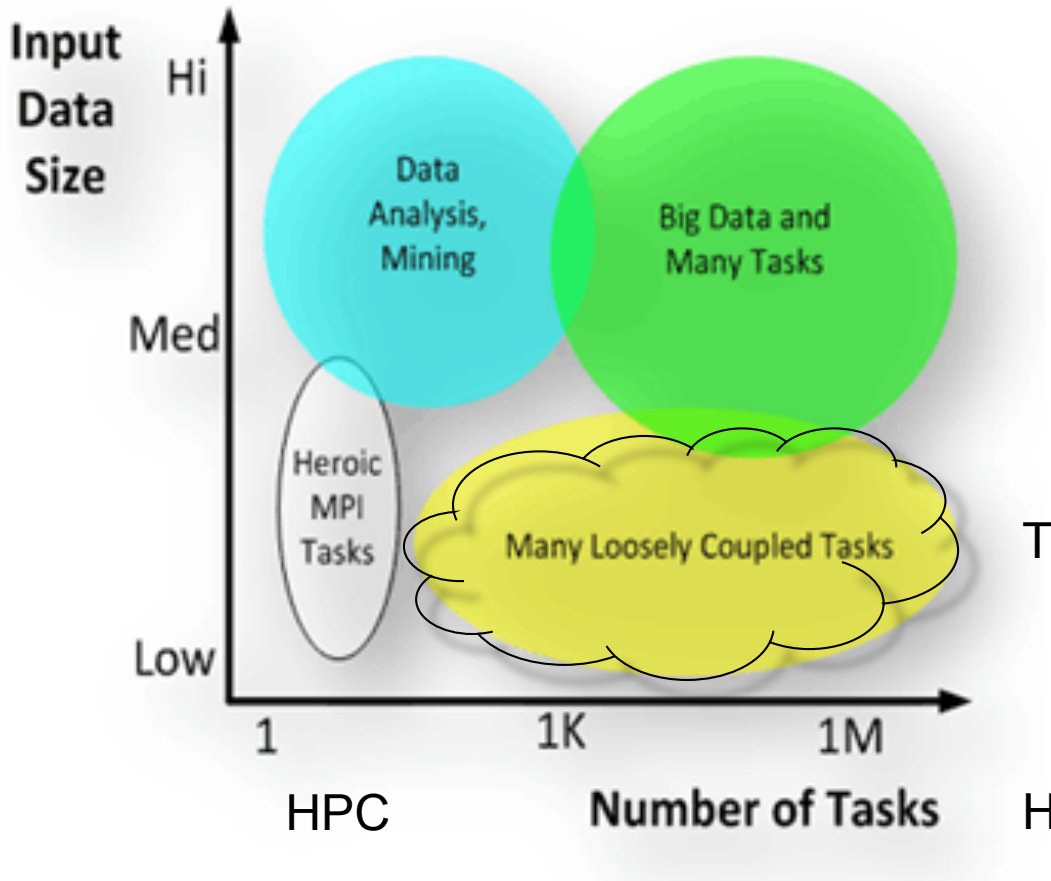
Tashi

http://wiki.apache.org/incubator/TashiProposal



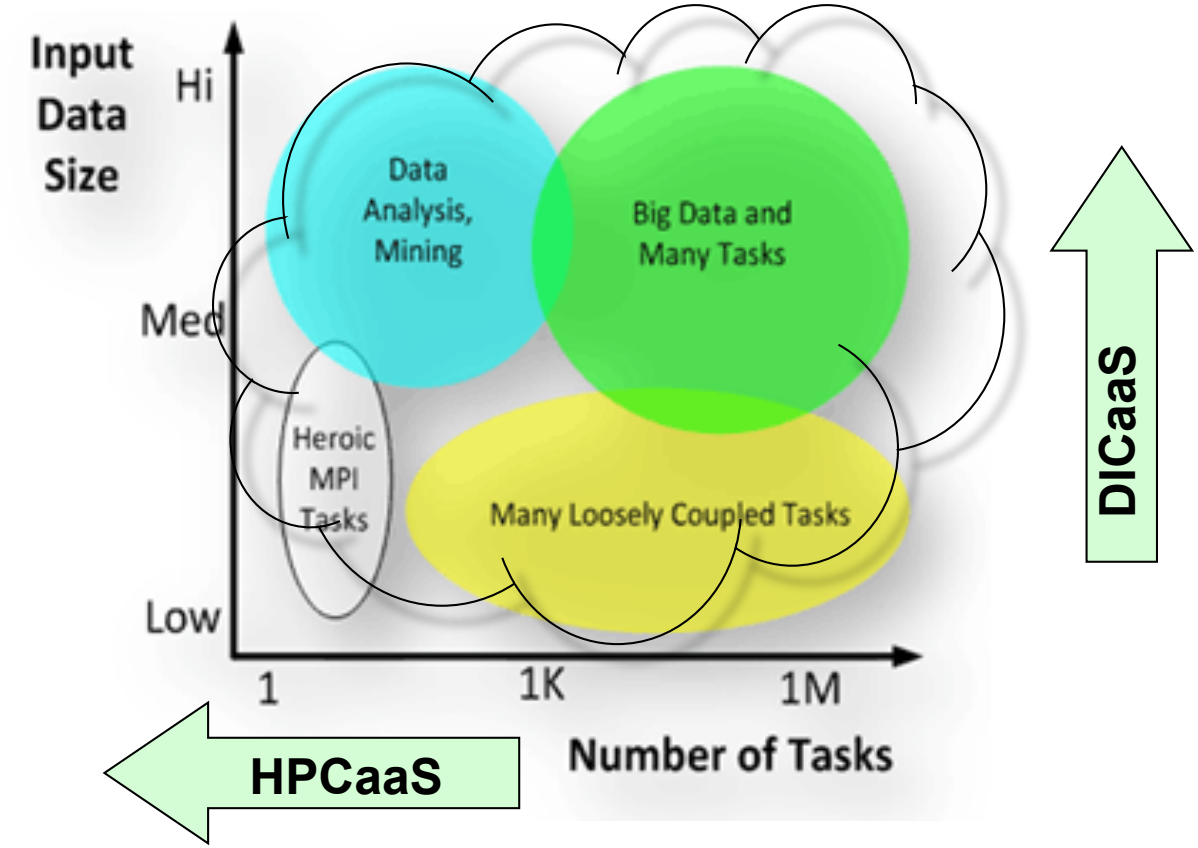
- Co-Scheduling of CPU and data

HPC vs. HTC vs. MTC (Many Task Computing)



Source: I.Foster

Extension of the Cloud Space



- **Cloud solutions for HPC and Data Intensive Computing**

HPCaaS

- **High Performance Computing as a Service**
- **Interesting fields for R&D in Open Cirrus**
  - Flexible platform services for HPC customers
  - Development of MPI services for clouds
  - Development of scheduling services for clouds
  - Management of software licenses

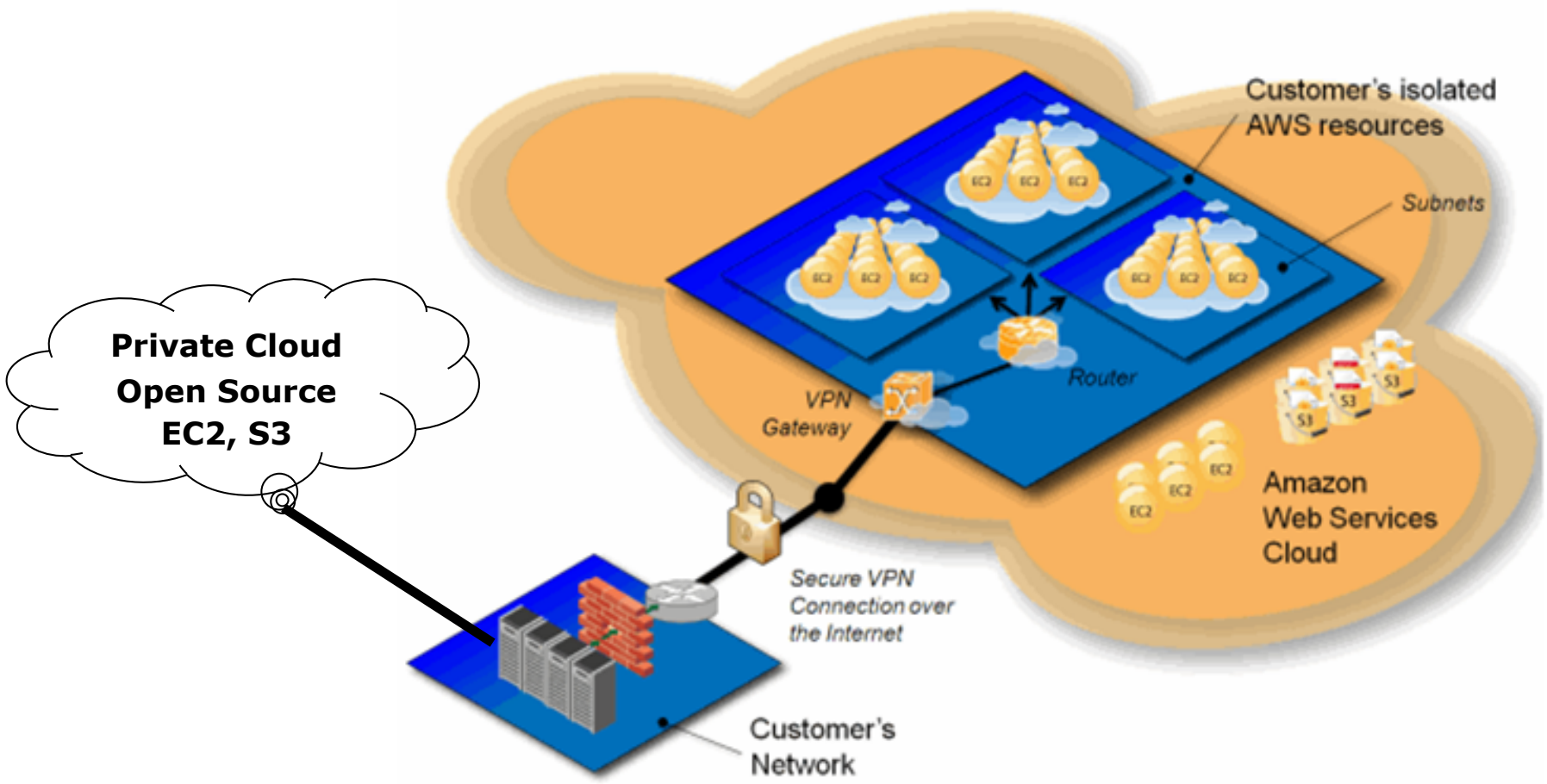
DICaaS

- **Data Intensive Computing as a Service**
- **Development of Big Data services**
- **Actual projects at KIT in this field:**
  - Data storage for LHC computing
  - Data storage for ITER (EUFORIA)
  - Project ANKA (synchrotron radiation source)
  - Activities in materials research
- **Cloud data archives**
  - Long-term data filing due to legal requirements

Cumulus: Private Cloud @ KIT

- Generate know-how to construct a private Cloud to support KIT
  - IT-Service supermarket
  - Self-service
  - Business models
- Work is based on open-source (PRS, Eucalyptus, Hadoop, …)
  - Performed in the framework of Open Cirrus
  - Currently PhD and diploma students
  - Needs further support (Sponsors are welcome)

Hybrid Clouds: Cloud Bursting



- **Transfer workloads and data transparently between clouds**
- **Example: Amazon Virtual Private Cloud**

Cloudy Issues

- **Public clouds are opaque**
  - What applications will work well in a cloud?
  - Data protection?
  - Security?
- **Many of the advantages offered by public clouds appear useful for "on premise" IT**
  - Self-service provisioning
  - Legacy support
  - Flexible resource allocation
- **What extensions or modifications are required to support a wider variety of services and applications?**
  - Data assimilation
  - Multiplayer gaming
  - Mobile devices

Notes from the Open-source Cloud

- **Private clouds are really hybrid clouds**
  - Users want private clouds to export the same APIs as the public clouds
- **In the enterprise, the storage model is key**
  - Scalable “blob” storage doesn’t quite fit the notion of “data file.”
- **Cloud federation is a policy mediation problem**
  - No good way to translate SLAs in a cloud allocation chain
  - “Cloud Bursting” will only work if SLAs are congruent
- **Customer SLAs allow applications to consider cost as first-class principle**
  - Buy the computational, network, and storage capabilities that are required
- **Further work has to be done especially in service research area**
  - Open-source is mostly dealing with technology related issues

Summary

- **Open-source cloud stack**
  - Standards are important to foster a cloud computing market
  - De-facto standards are around (e.g. Amazon EC2, S3)
  - Standards are evolving (OCCI)
  - Federation of private and public clouds (Cloud bursting)
- **Open Cirrus project offers interesting R&D opportunities**
  - Cloud systems development
  - Cloud application development
  - Accepting research proposals
- **Generate know-how to construct a private Cloud @ KIT (Cumulus)**
  - IT-Service supermarket
  - Self-service
  - Business models


©johnlund.com

Access to Open Cirrus Sites

- **Project PIs apply to each site separately**
- **Contact names, email addresses, and web links for applications to each site will be available on the OpenCirrus Web site**
  - http://opencirrus.org
- **Each OpenCirrus site decides which users and projects get access to its site**