CENTRE DE SUPERCOMPUTACIO DE CATALUNYA

CESCA

Repositorios en la nube

Ricard de la Vega
Jefe del Servicio de Portales y Repositorios
Centre de Supercomputació de Catalunya

4as Jornadas OS-Repositorios
Barcelona, 3-5 marzo de 2010

ANELLA CIENTÍFICA CATNIX TDX RACO RECERCAT mdx JOCS TAPC TSIUC TERAFLOP

Centre de Supercomputació de Catalunya

- ✓ Consorcio público
- ✓ Creado en 1991
- ✓ Formado por:
  - Generalitat de Catalunya
  - Fundació Catalana per a la Recerca i la Innovació
  - 9 universidades catalanas
  - Consejo Superior de Investigaciones Científicas
- ✓ Anella Científica creada en 1993

ANELLA CIENTÍFICA

Nuestros servicios

Comunicaciones

ANELLA CIENTÍFICA

Red IRIS Node Catalunya

CATNIX

Portales y Repositorios

TDX

RECERCAT

RACO

RECYT

mdx

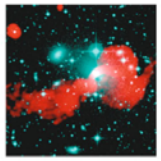
padicat

MDC

CALAIX

Cálculo y Archivo

CAP

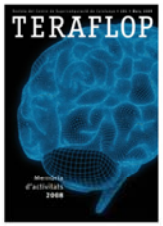


SED

SDF

AUC

Promoción



JOCS

TAPC

TSIUC

Operaciones y Seguridad

SAH
EC-UR
ERIAC
S24x7

Agenda

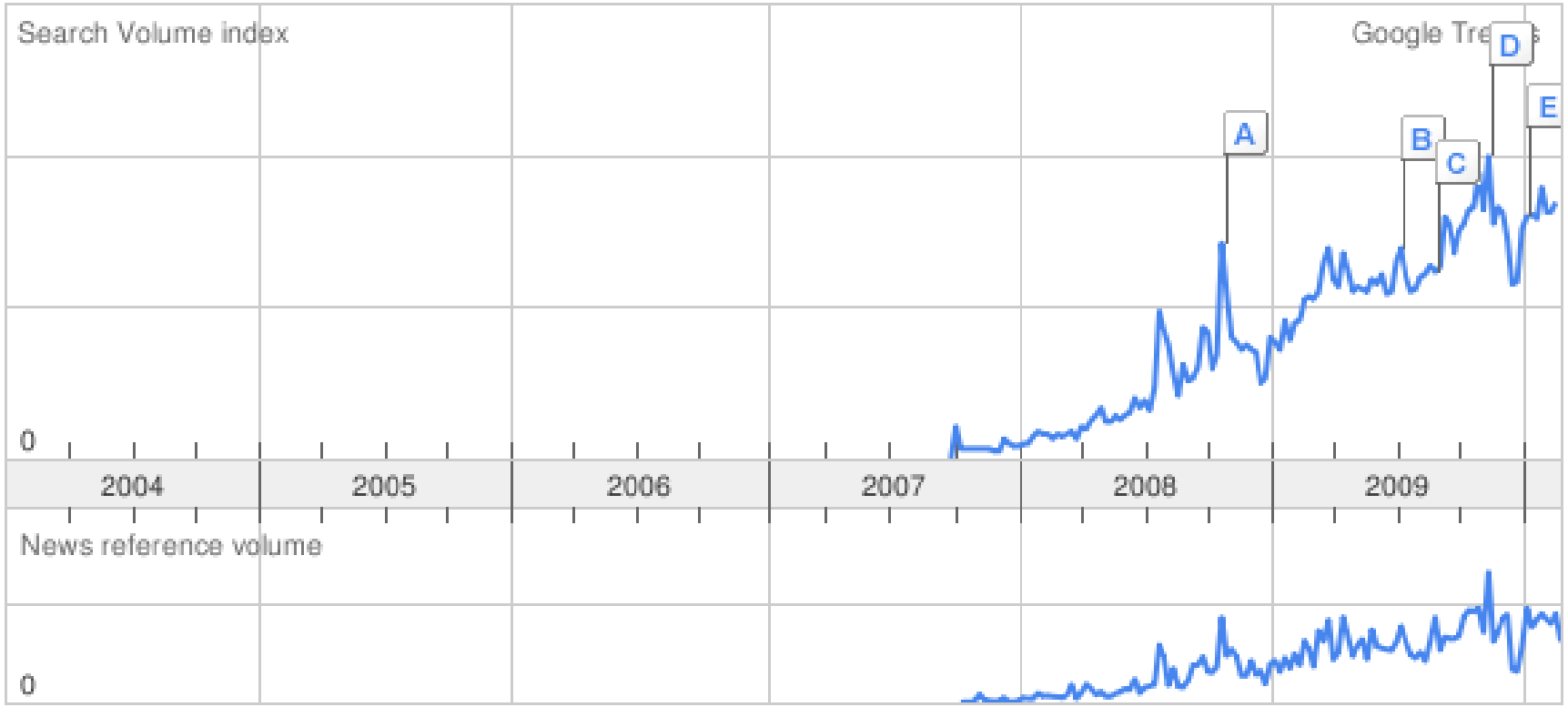
- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Tendencia en las TIC

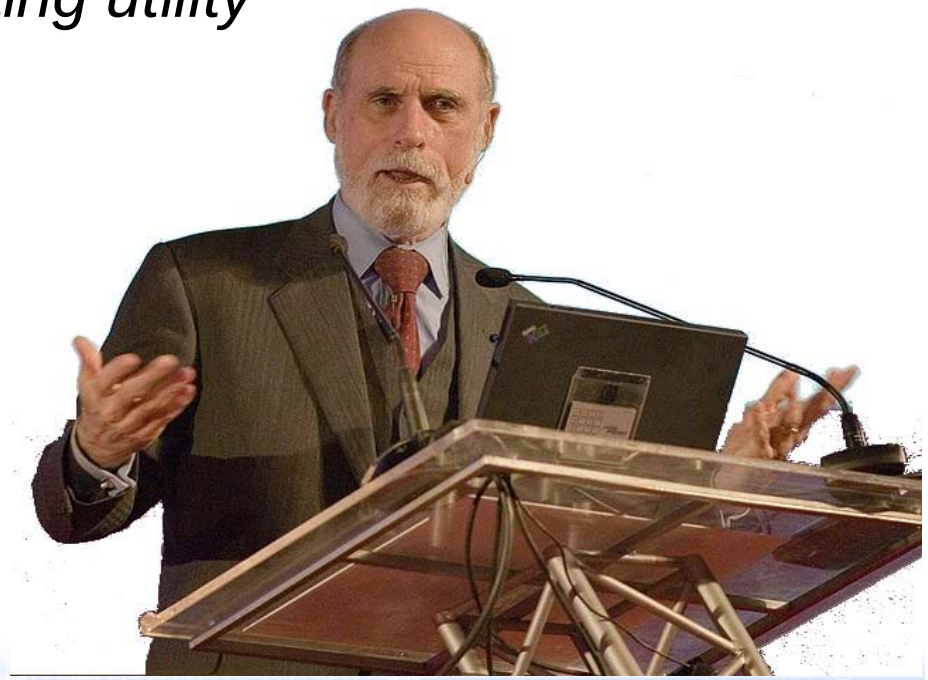
● cloud computing



“Cuanto más cambia una cosa...”

- ✓ “Cuanto más cambia una cosa, más se convierte en lo mismo. Fijémonos en el *cloud computing*. En cierto modo, se trata de una extensión natural del *time-sharing,* inventado en los años setenta. De hecho, por aquel entonces ya se oía hablar del término *computing utility*”

Vinton Cerf



Mainframes...


VAX 8600


Controladora discos HSC50


Terminal Digital VT220


Impresora LA-36

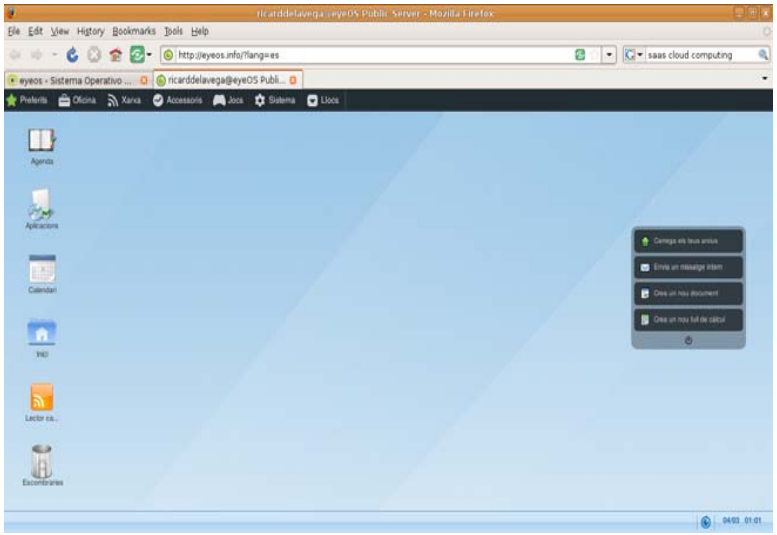

TA78

Fuente: *http://fib.upc.edu/retroinformatica/exposicio/ordinadors.html*

Clouds...

Centros de datos

El terminal es el navegador



Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Definiciones y conceptos clave

"The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing.

We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public."

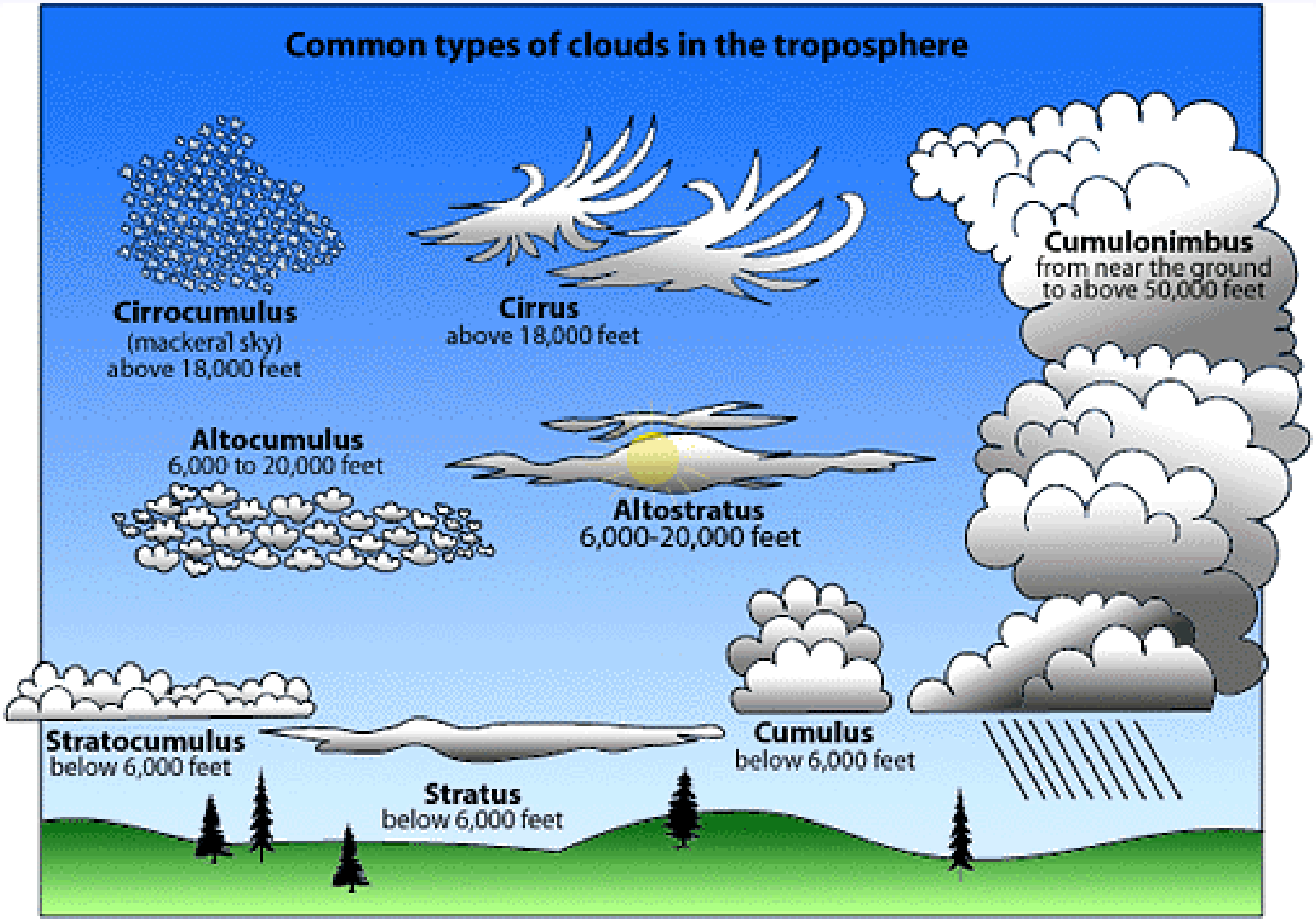
Conceptos: "as a Service", pago por uso, virtualización

Fuente: *http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html*

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Tipos de *cloud*



Fuente: *http://www.ucar.edu/educ_outreach/webweather/cloud3.html*

“as a service”

- ✓ Software como un servicio (SaaS)
  - Aplicaciones de Google (Gmail, Calendar...)
  - Salesforce.com
- ✓ Plataforma como un servicio (PaaS)
  - Imagen de Xen con SO, Apache, MySQL y aplicación
  - Google App Engine
  - Microsoft Azure
- ✓ Infraestructura como un servicio (IaaS)
  - Amazon Web Services
    - EC2 para computo
    - S3 para almacenamiento

Pública, privada...

- ✓ Una **nube pública** es el hardware y software de un centro de datos ofrecido en la modalidad de “pago por uso”.
- ✓ Una **nube privada** es el hardware y software de un centro de datos de la propia entidad.
- ✓ Una **nube híbrida** combina los dos modelos anteriores.
- ✓ Inconvenientes de la nube pública:
  - Privacidad y protección de datos

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Participantes y beneficios

- ✓ **Transferencia del riesgo** a los proveedores del cloud
  - Service Level Agreement (SLA)
- ✓ Proveedores del cloud (hardware y software de base)
  - Reducción de los costes de operación y amortizaciones
  - Especialización y economías de escala
- ✓ Usuarios de la nube, que a su vez, son proveedores de servicios (SaaS), como los **repositorios**
  - Elasticidad en el aprovisionamiento de recursos
  - Sin sobredimensionamiento ni infradimensionamiento (picos)
- ✓ Usuarios finales, como los investigadores, etc.
  - Para ellos la nube es transparente

Como el *car sharing*

Carsharing avancar

coches por horas
muy cerca de ti


Seat Ibiza
o equivalent


Renault Mégane
o equivalent


Honda Civic Hybrid


Renault Mégane
Grand Scénic


Opel Vivaro

Carsharing es...

- disponer de un **coche** cuando lo necesitas, **por horas** o por días
- inmediato, flexible, y muy **cerca de ti**
- a precios económicos, **todo incluido** y sin sorpresas
- respetuoso con el **medio ambiente**

...muy fácil

Date de alta con una **cuota anual** a partir de **30€**, y recibirás la **tarjeta Avancar**, que te da acceso a nuestros coches:

1

Reserva

- por Internet o teléfono **24 horas, 7 días**
- el coche que te conviene, en el aparcamiento más cercano
- para el tiempo que necesites, desde **1 hora** hasta varios días

2

Abre el coche

- encontrarás el coche en su **plaza reservada**
- ábrelo con tu **tarjeta Avancar**
- y el coche es tuyo...

3

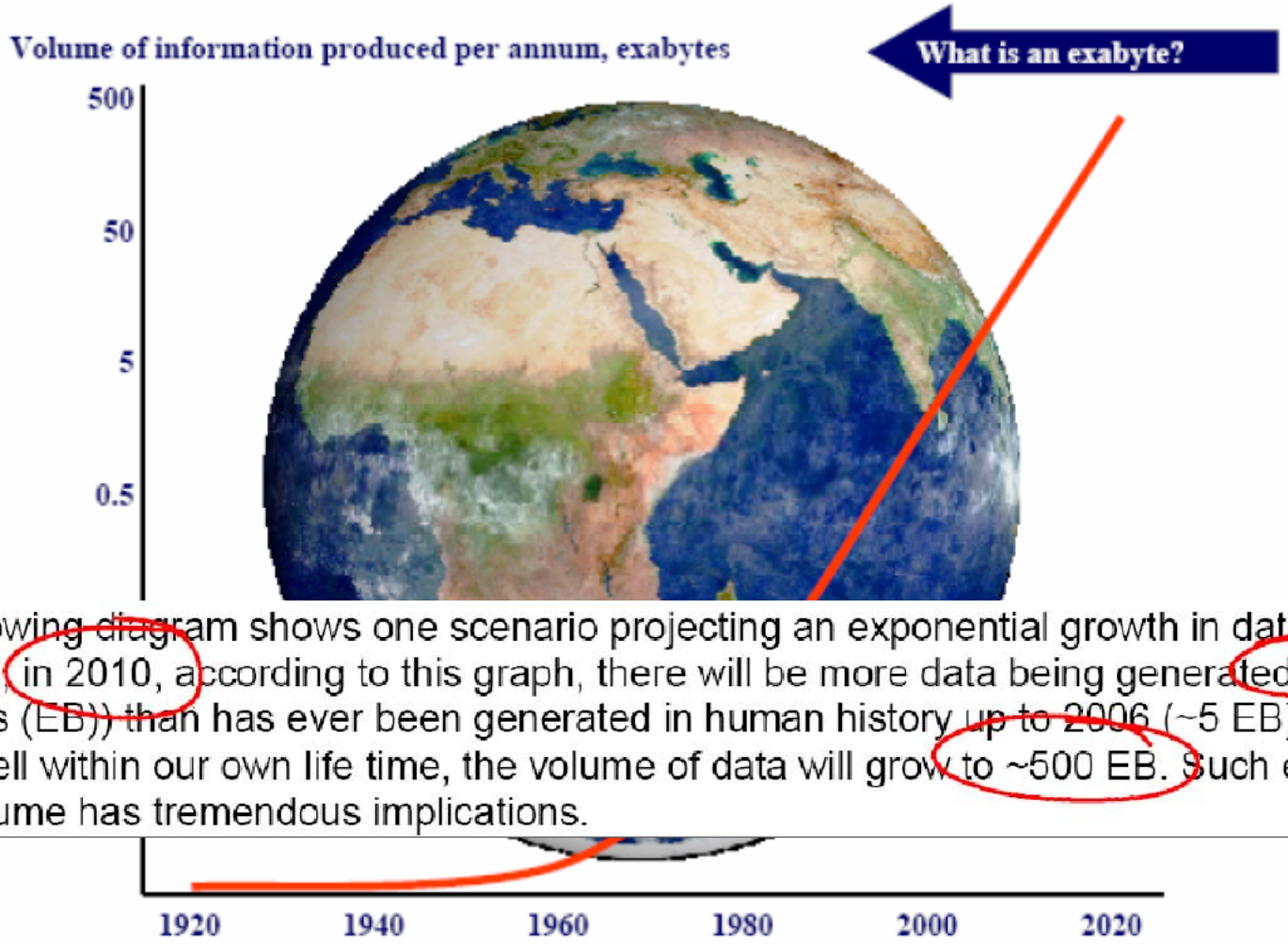
Conduce

- pagas desde **3.15 €/hora** + **0.24 €/km**, combustible incluido
- con **seguro a todo riesgo** y **asistencia 24h**
- al final del viaje, dejas el coche en su **plaza reservada**

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Data deluge



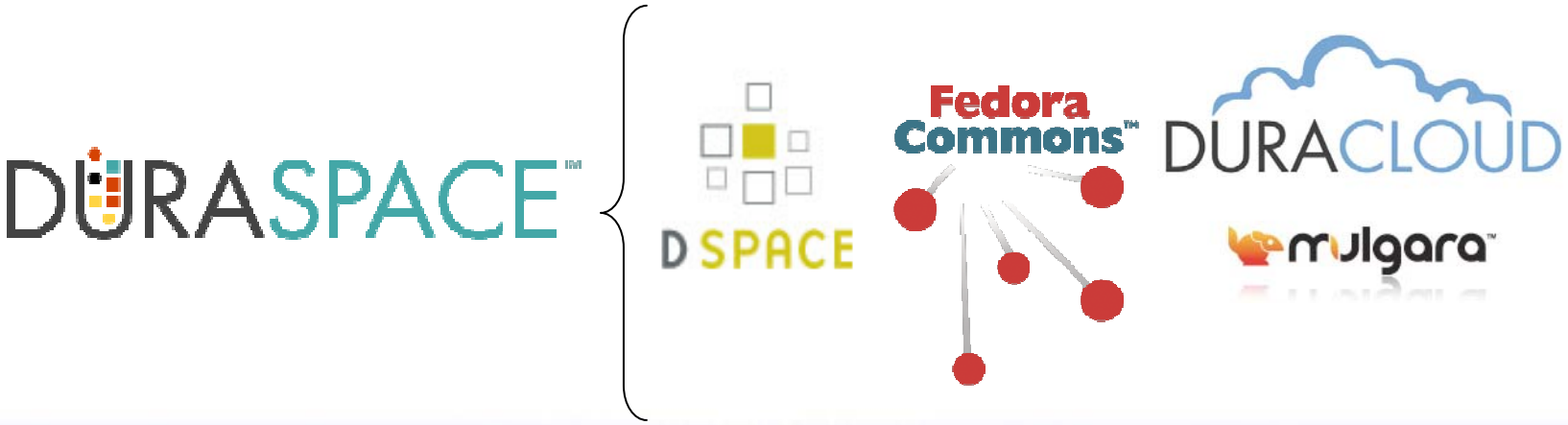
The following diagram shows one scenario projecting an exponential growth in data. For example, in 2010, according to this graph, there will be more data being generated (~50 exabytes (EB)) than has ever been generated in human history up to 2006 (~5 EB). By the year 2020, well within our own life time, the volume of data will grow to ~500 EB. Such explosion of data volume has tremendous implications.

Agenda

- ✓ Duracloud
  - DURASPACE
  - Objetivos
  - Servicios
  - Arquitectura
  - Pilotos
  - Roadmap

DURASPACE

- ✓ DuraCloud es de DURASPACE
  - Non-profit-org que da soporte a las comunidades Dspace y Fedora.
  - Innovación:
    - Pensar más allá de las plataformas actuales.
    - Nuevas estrategias para el acceso y la preservación de contenidos digitales.

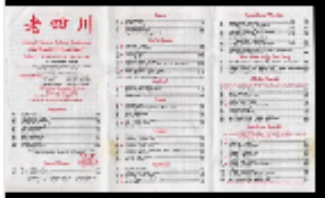


Objetivos

- ✓ Soporte a la preservación
  - Replicación de contenido, auditoría (*checksums*), reparación
- ✓ Federación de repositorios y ciberinfrastructura
  - Enlaces entre datos almacenados (*linked data*)
- ✓ Colecciones compartidas
  - Acceso vía un motor JPEG2000 a imágenes almacenadas
- ✓ Data mining
  - Grandes trabajos de computación con los datos almacenados

Servicios básicos

DuraCloud - basics



Chinese Menu
of Service Options

Replicate to multiple storage providers

Replicate to multiple geographic areas

Monitor and audit digital assets

Compute services in cloud next to content

Hosted by DuraSpace not-for-profit org

Partnerships with cloud providers

"Pay for use" for services and storage

Available to run internally- open source

Fuente: *http://www.duraspace.org/documents/DuraCloudNDIIPPJune09.ppt*

Servicios adicionales

Additional services

- Other DuraSpace-provided services on top of content stored in the cloud
  - Search
  - Aggregation
  - Streaming
  - Migration
  - Hosting repositories

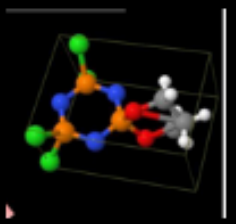
Fuente: *http://www.duraspace.org/documents/DuraCloudNDIIPPJune09.ppt*

Casos de uso con almacenamiento en la nube

Use Cases: DuraCloud with Cloud Storage

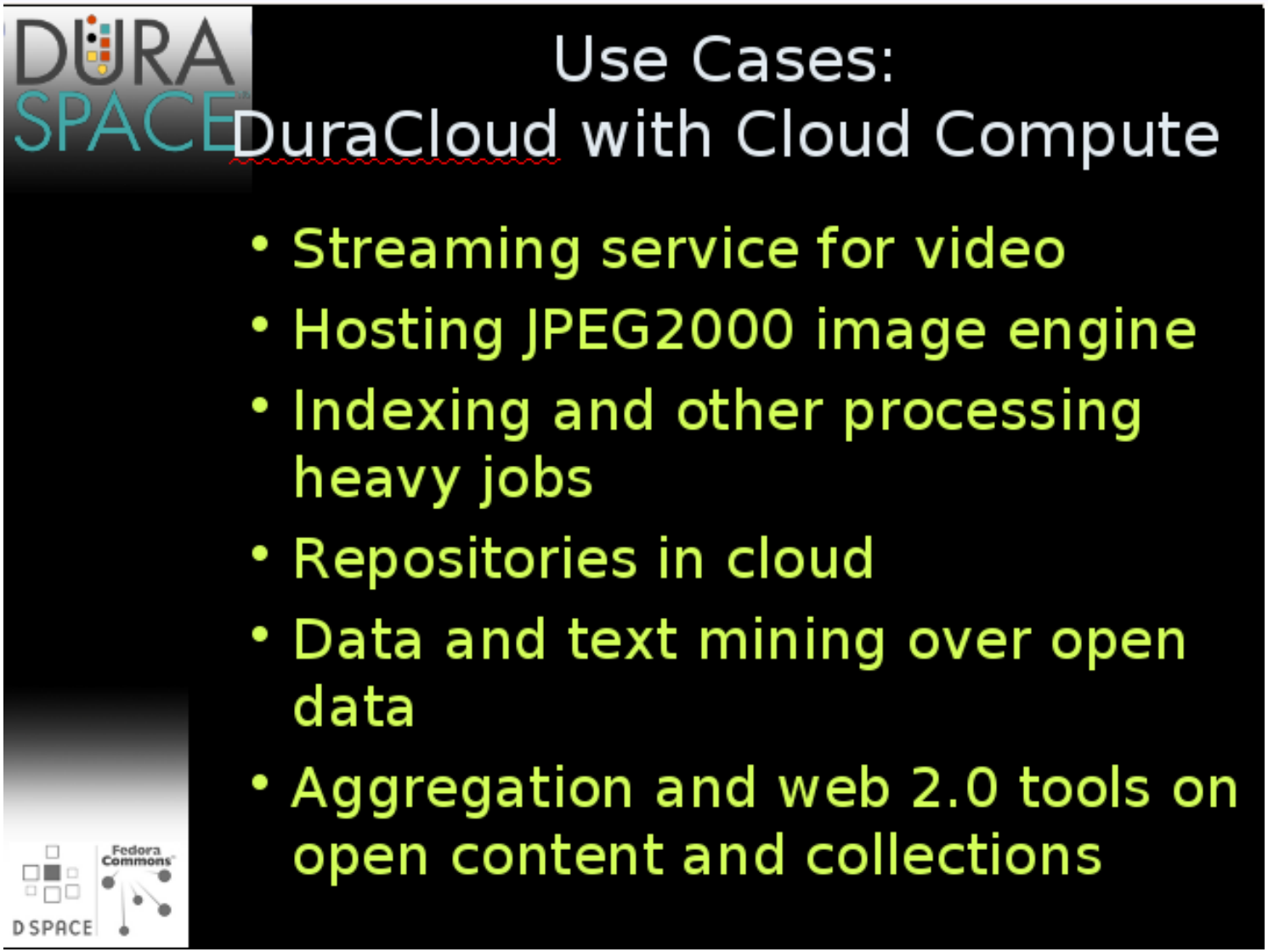
- Online backup for text, images, datasets, video, audio
- Enable preservation via multiple copies, geographies, administrations
- Elastic provisioning of temporary or permanent storage for projects or jobs





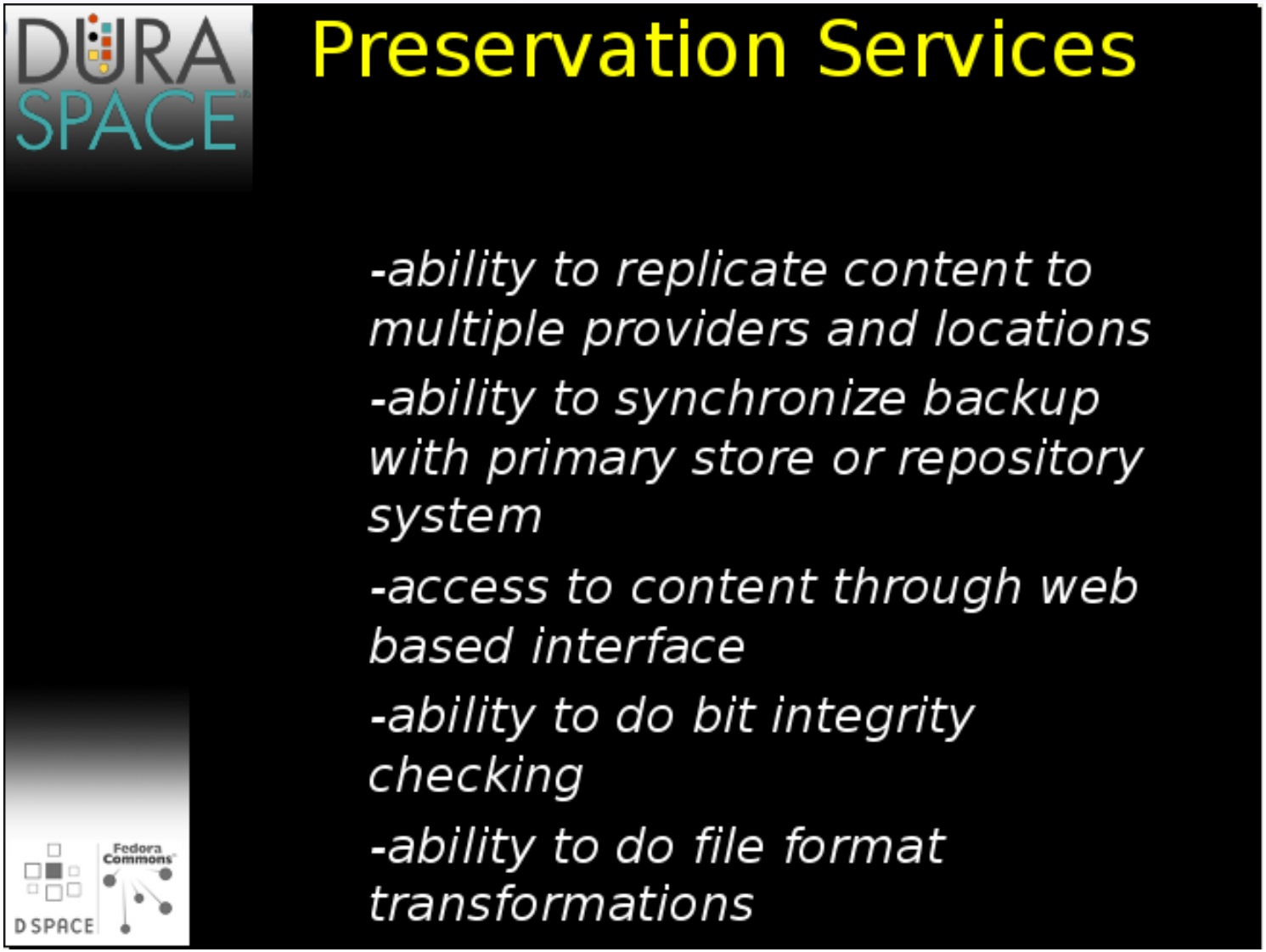

Fuente: *http://www.duraspace.org/documents/DuraCloudNDIIPPJune09.ppt*

Casos de uso con computación en la nube



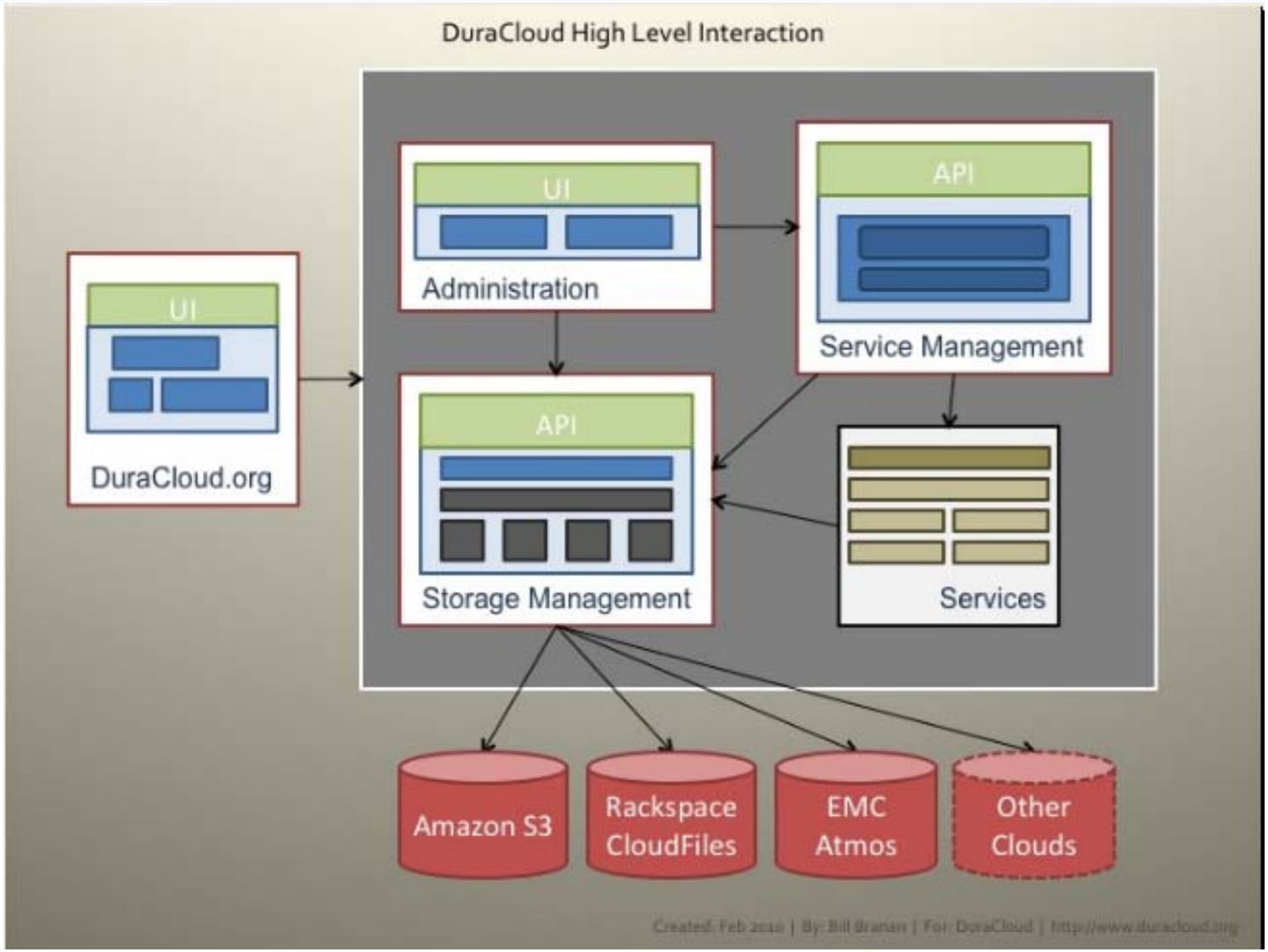
Fuente: *http://www.duraspace.org/documents/DuraCloudNDIIPPJune09.ppt*

Servicios de preservación



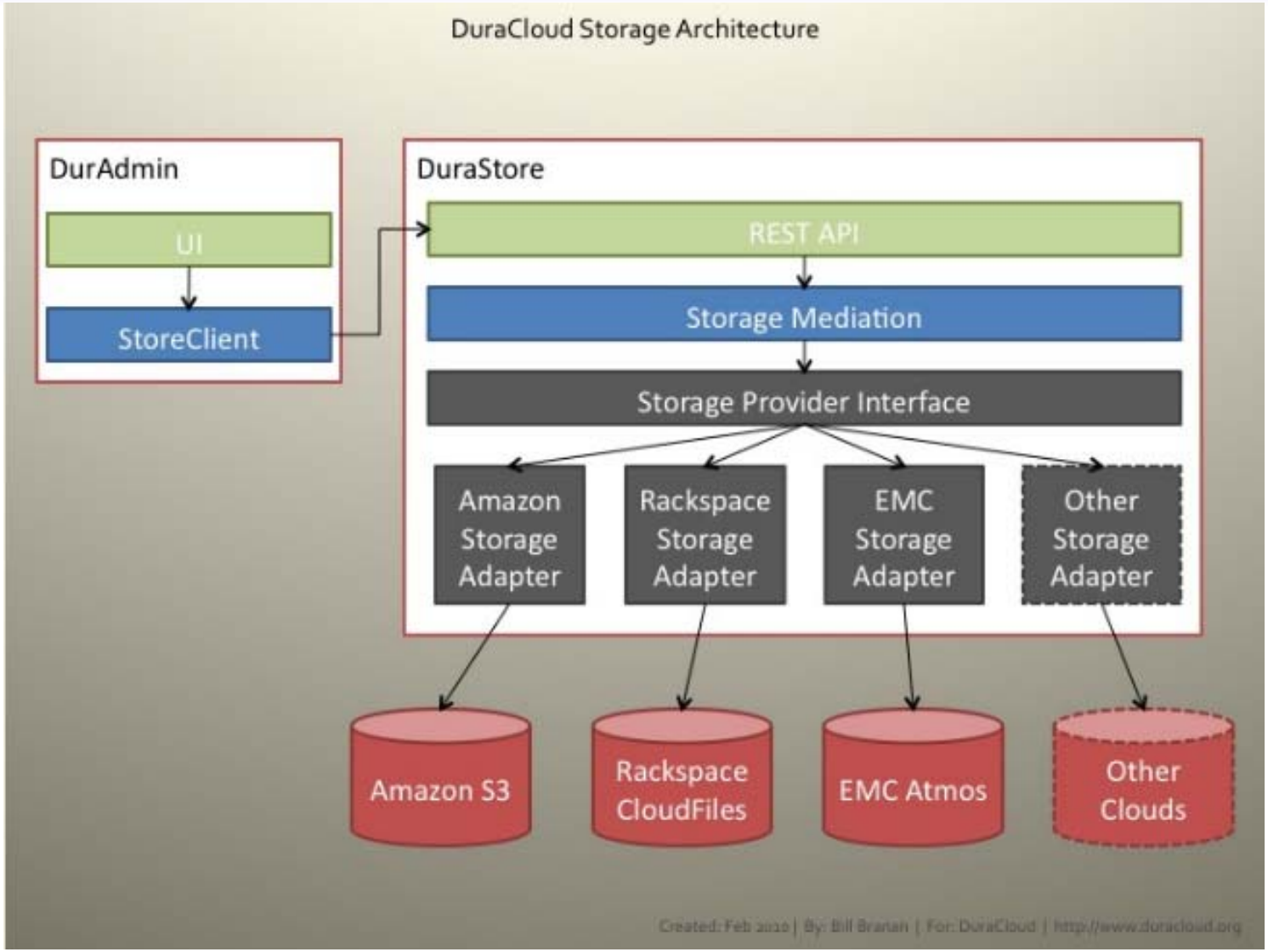
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Arquitectura



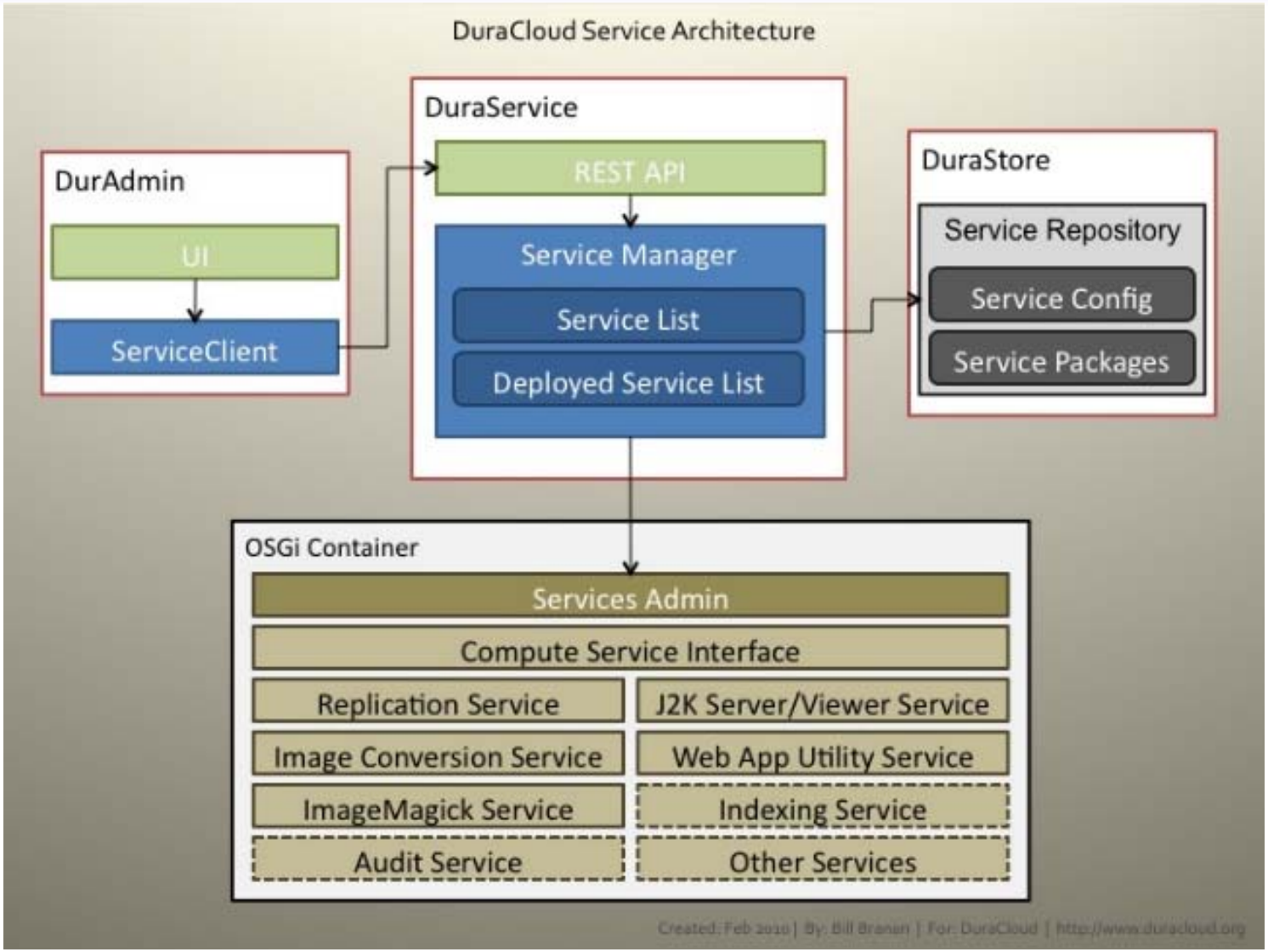
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Arquitectura



Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Arquitectura



Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de casos de uso

- ✓ Quieren...
  - Introducción de gran cantidad de contenidos digitales
  - Replicar a múltiples plataformas de *cloud*
  - Gestionar esta replicación y monitorizarla
  - Desarrollar servicios

- ✓ Proveedores de *cloud*
  - Amazon
  - EMC
  - Rackspace
  - Sun?
  - Microsoft?

- ✓ *Partners* iniciales
  - New York Public Library
  - Biodiversity Heritage Library
  - WGBH Media Library and Archives

Pilotos de casos de uso

NYPL pilot

Digital Gallery Collection

Use case: back up online preservation copy to Fedora, file format transformation

-back up copy all TIFF images (10 TB data)

-transformation from Tiff to JPEG 2000 using Imagemagick

-run J2k image server in cloud

-Push JPEG 2000 back into Fedora Repository

Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de casos de uso

DURA SPACE

BHL pilot

BioDiversity Heritage Library

Use case: Find the best cost competitive solution for keeping multiple copies in multiple geographies, easily accessible.

- -back up copy entire corpus (40 TB data-JPEG, Tiff
- -have multiple copies including Europe
- -Run J2K image server in cloud

Fedora Commons

DSPACE

Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de casos de uso

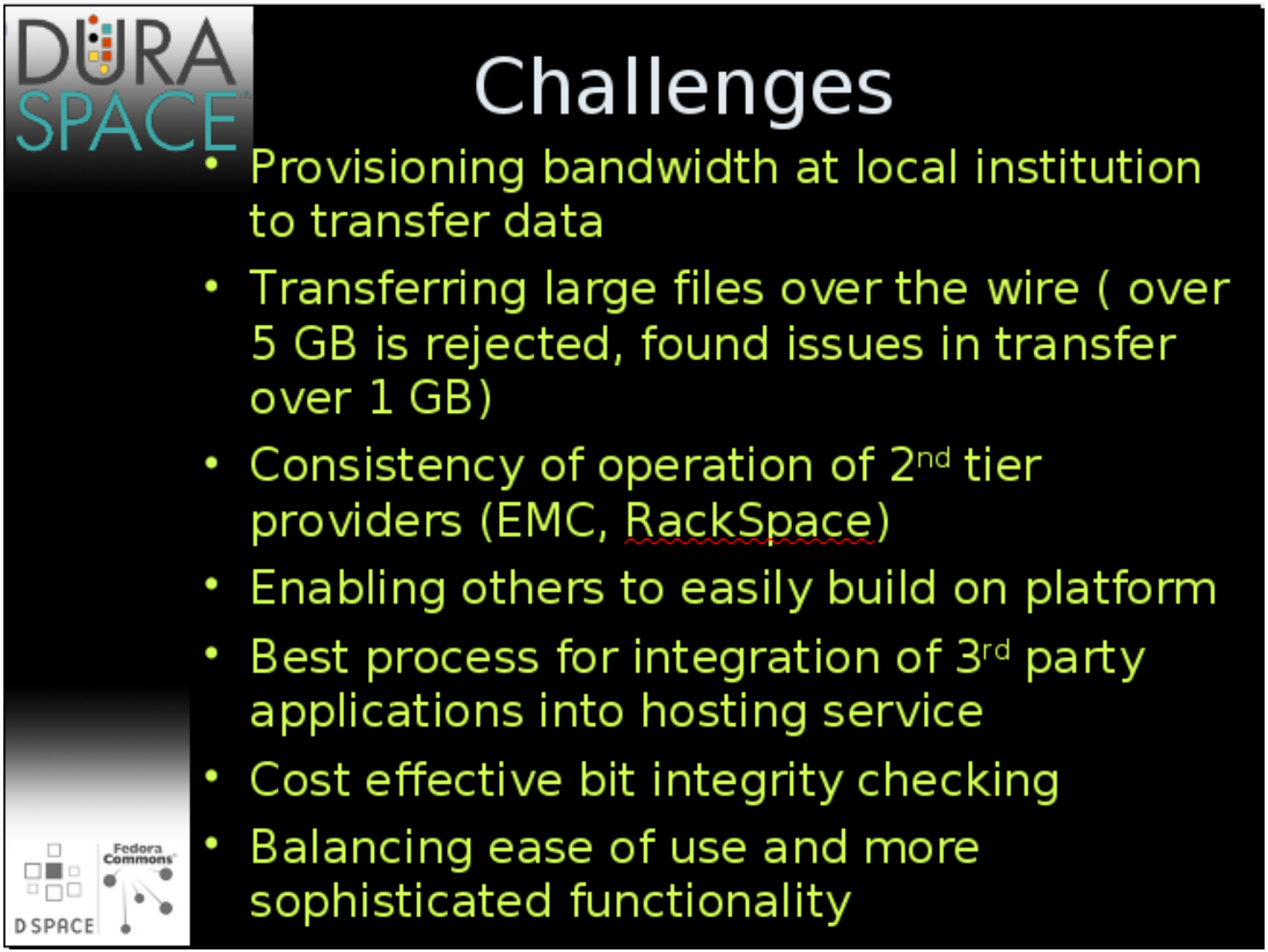
WGBH Media Library and Archives

Use case: Provide backup preservation for video files from repository and other sources, and create derivative files for access and streaming.

- Archive large video files
- Provide public access to streaming versions
- Transcode files in cloud
- Edit files where appropriate to sell clips
- Give third party access to cloud store for processing and access

Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de casos de uso, retos



Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de *beta testers*

DuraCloud Pilot

Pilot Requirements

We are looking for volunteer organizations to participate in the DuraCloud beta test program that are interested in experimenting with and ultimately using DuraCloud for preservation support. The pilot will run for three months in late spring (exact dates to be determined). During the pilot period you will be required to do the following:

1. Set up a duracloud account
2. Transfer your data from your primary site into DuraCloud
3. Use the DuraCloud web application to manage your data and sets of services
4. Provide feedback during the 3 month period through the following media: wiki, jira, skype chat, personal discussion, and online survey

Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Pilotos de *beta testers*

Limitations

Data transfer into DuraCloud will be limited to 500 GB of data. The pilot will be no cost to the organization, other than personnel time to participate. After the pilot has concluded the organization will have the option to do one of the following:

1. Keep the DuraCloud account and transfer to production service
2. Remove/delete all data from DuraCloud and cloud providers

Disclaimer During Beta Testing Period

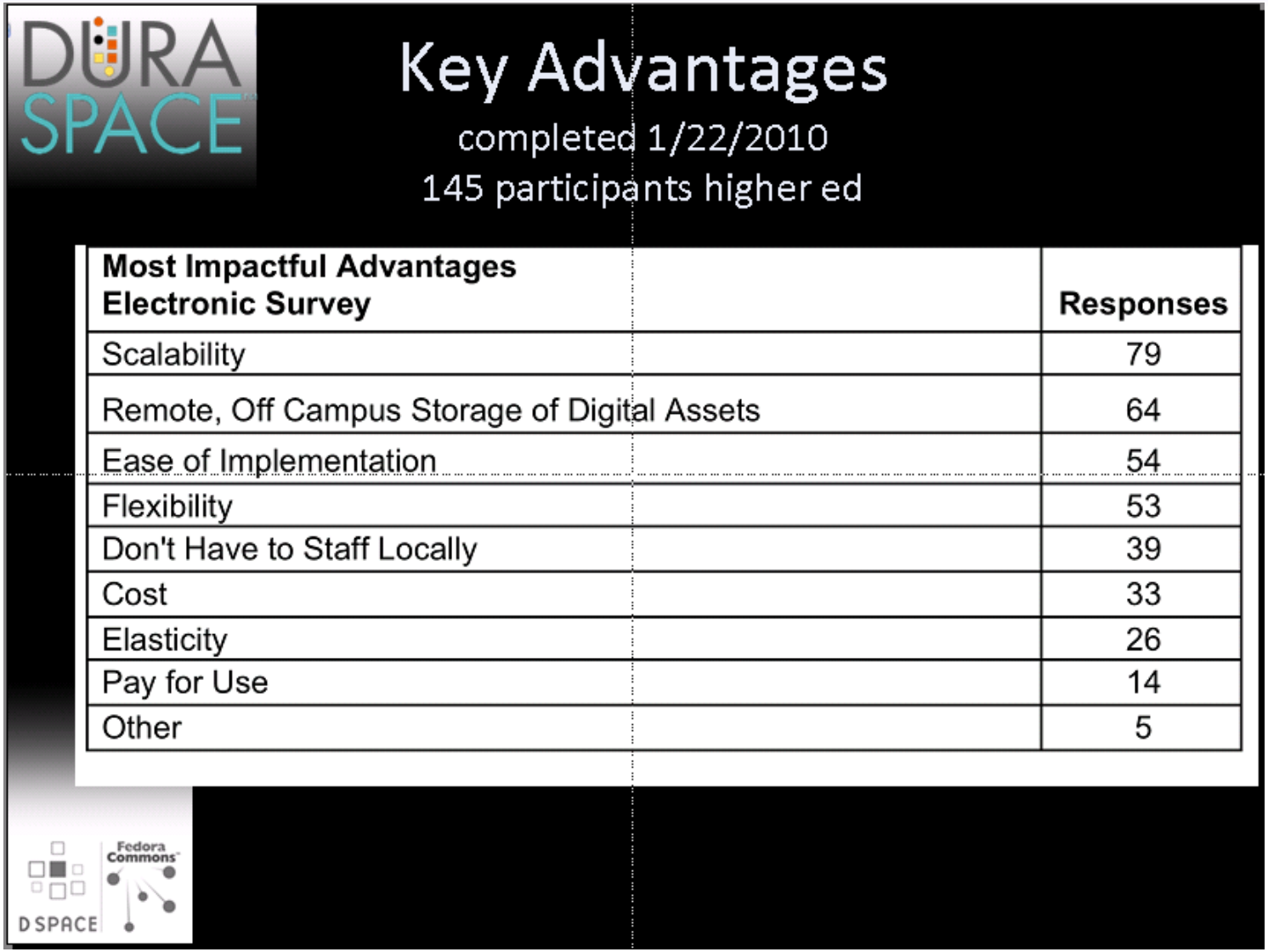
- DuraCloud is not responsible for data damaged or lost during the course of the pilot.
- DuraCloud is not liable for breach of service by any of its cloud provider partners during the course of the pilot.
- DuraSpace does not take ownership of any content transferred into DuraCloud.

DuraCloud Pilot Form

Please complete and submit the form below to be considered as a DuraCloud pilot applicant.

Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Encuesta



Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Encuesta

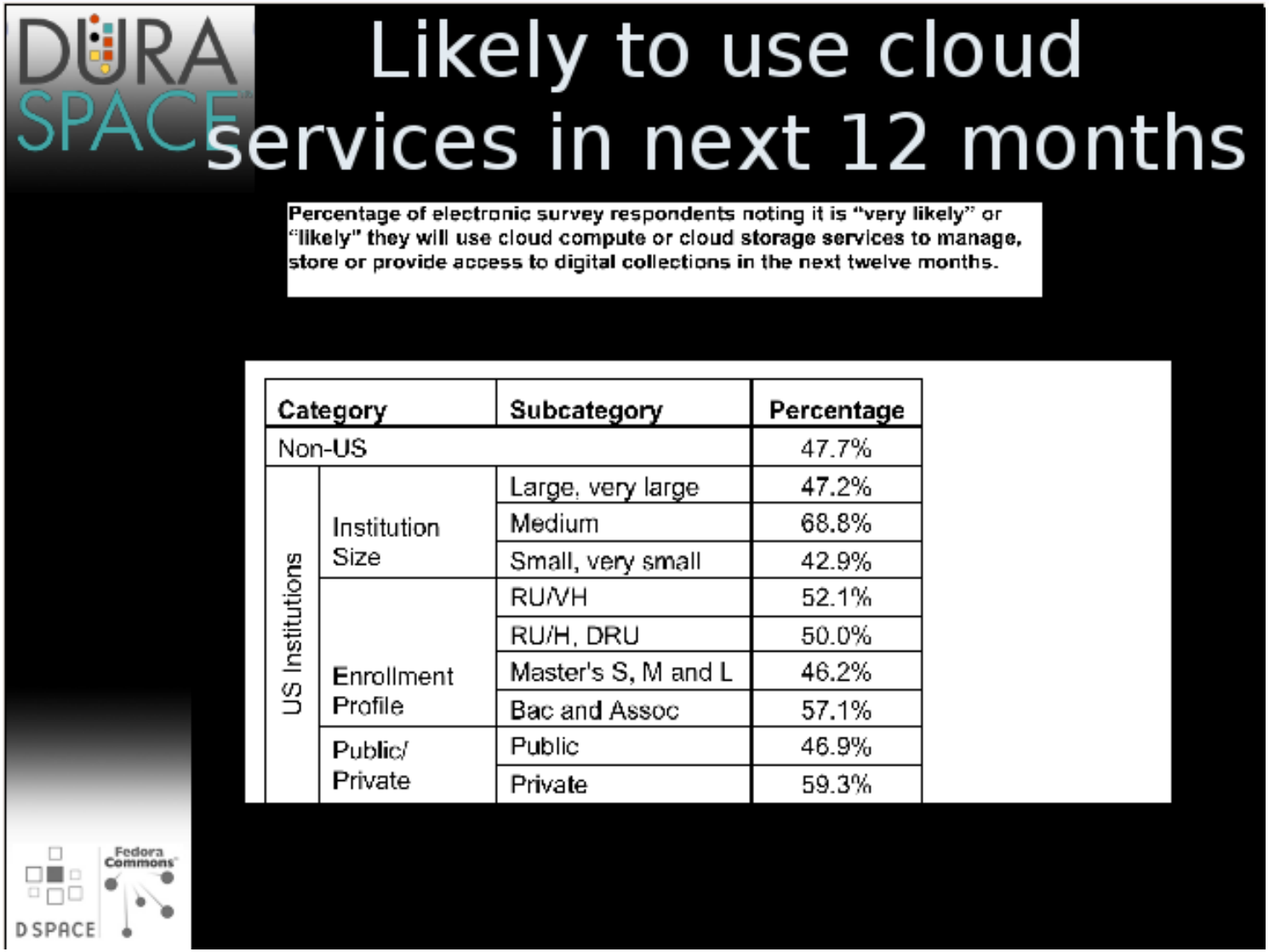
Key Challenges

completed 1/22/2010
145 participants higher ed

Key Challenges Electronic Survey	Responses
Trusting Third Party to Manage Critical Assets	64
Long-term Reliability of Solution	52
Data Security	51
Performance and Bandwidth Concerns	37
Loss of Control	34
Administrative Burden of SLAs	17
Transparency of Solution	16
Concerns about Data Lock-in	16
Less Customizable	10
Other	12

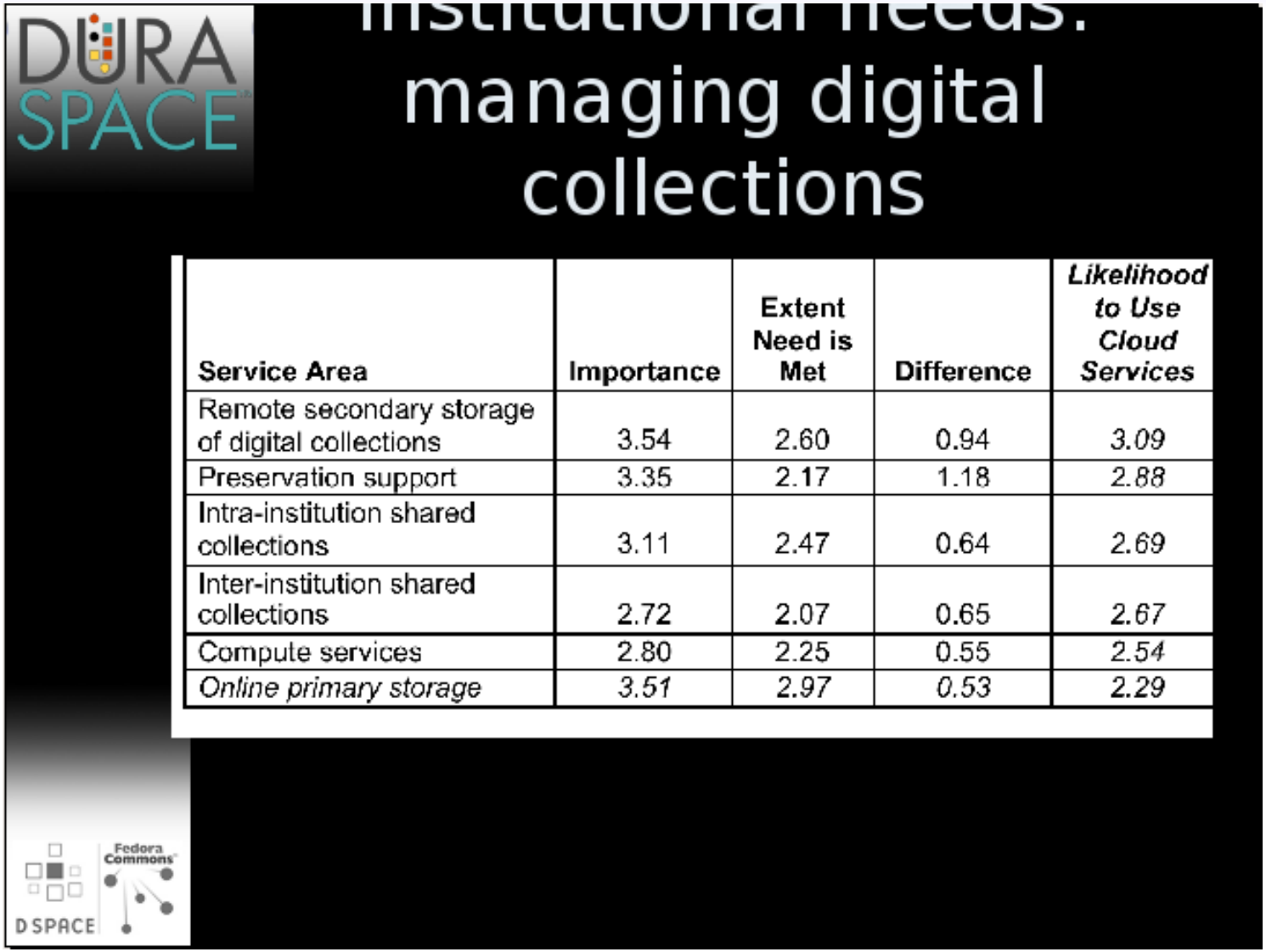
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Encuesta



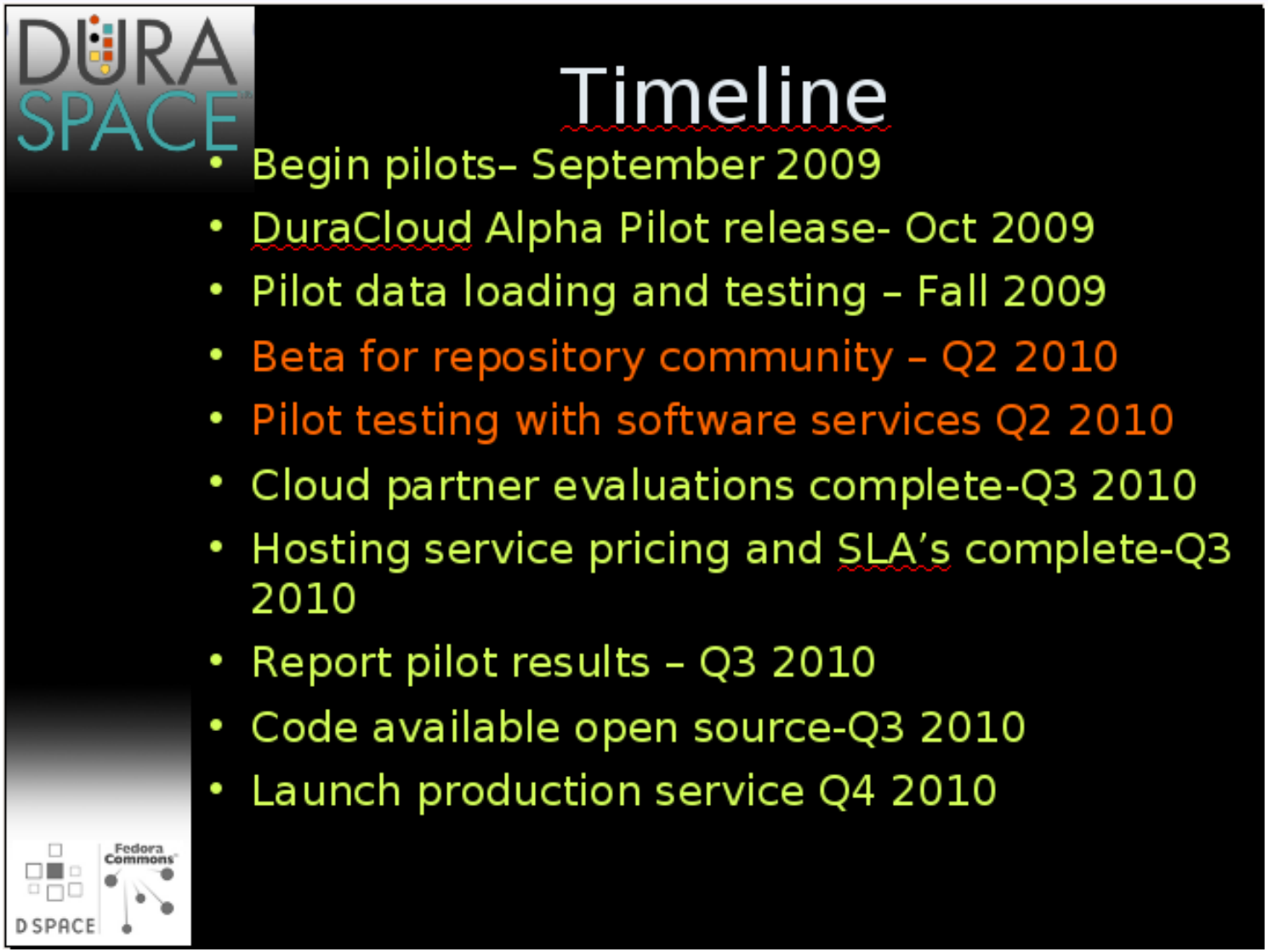
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Encuesta



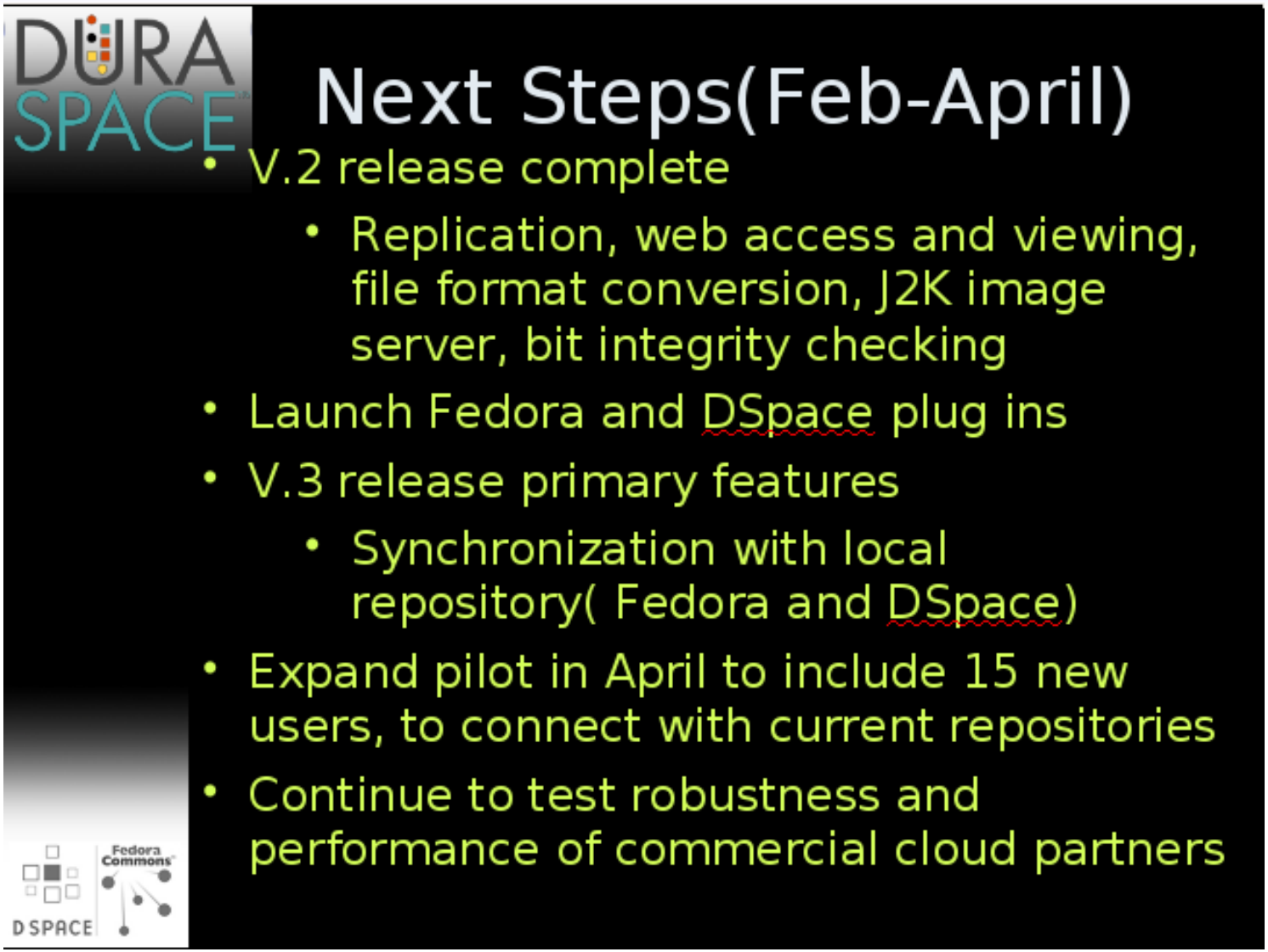
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Roadmap



Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Roadmap



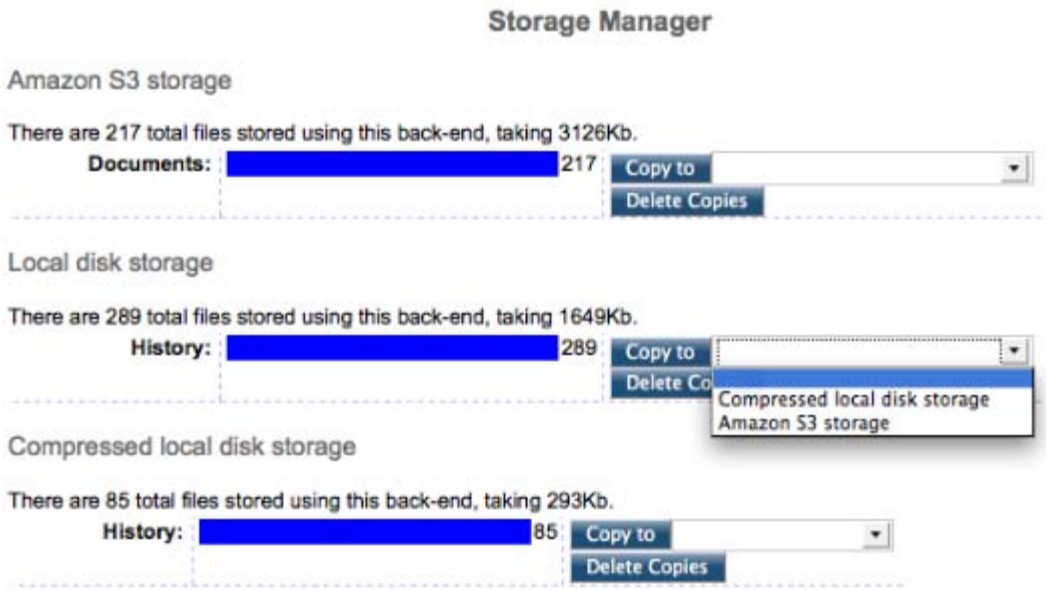
Fuente: *http://www.slideshare.net/eduserv/duracloud-open-technologies-and-services-for-managing-durable-data-in-the-cloud*

Agenda

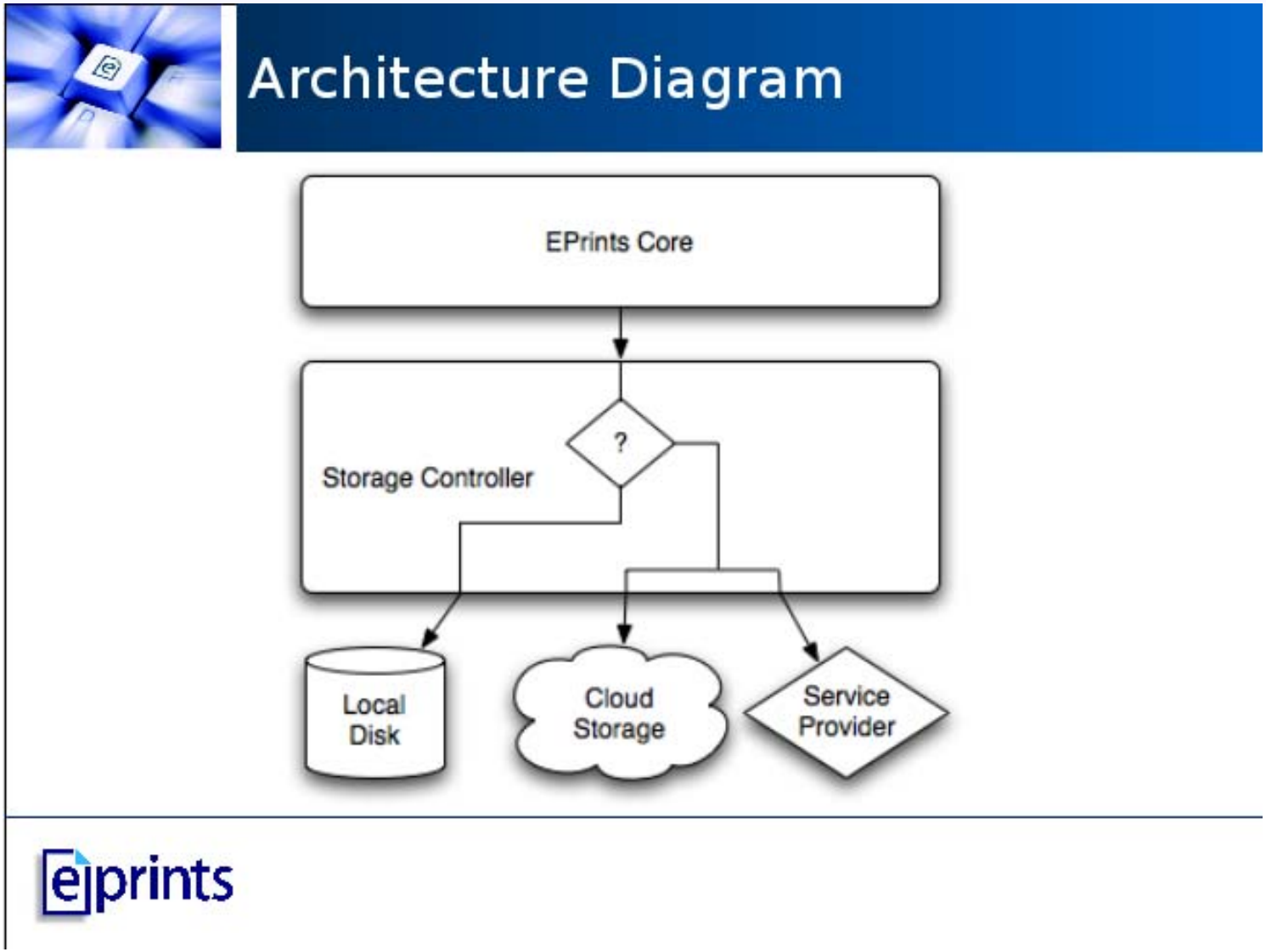
- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Eprints Cloud Capabilities

- ✓ Los datos se pueden almacenar en:
  - En disco local o cabinas de discos (SAN, NAS)
  - En el *cloud*
- ✓ Mediante **storage controller** se puede elegir a que tipo de disco van los documentos (archivo XML de configuración).



Arquitectura



Fuente: *http://www.slideshare.net/lescarr/eprints-and-the-cloud*

Plug-ins de almacenamiento en *cloud* en el API

Storage Plug-ins

- Local
- NFS
- Amazon S3
- Sun Cloud Storage Service
- Microsoft Azure
- Any others based on the S3 API…. (the last 3 all are)

- 5 Call API (about 30mins to write a plug-in)

eprints

Fuente: *http://www.slideshare.net/lescarr/eprints-and-the-cloud*

Objetivos

Our Development Vision

- Empower the Community with a simple API
  - API in 3.2
- Give the community a platform to test their code
  - Use the Cloud!
- Give the community a distribution mechanism
  - The EPrints Bazaar (beta)

Fuente: *http://www.slideshare.net/lescarr/eprints-and-the-cloud*

EPrints Bazaar

EPrints Bazaar

- Similar in concept to Apple's App Store
- Every install of EPrints will have access to the Bazaar
- Single click install/uninstall of plug-ins
- EPrints Services Approved Plug-ins
  - Enterprise support for limited 3rd party plug-ins

eprints

Fuente: *http://www.slideshare.net/lescarr/eprints-and-the-cloud*

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Reflexión antes de las conclusiones...

- ✓ En este contexto (en las nubes), se podría decir que los repositorios cooperativos (TDR, RECERCAT...) están en un *private cloud*, puesto que las instituciones que los coordinan, CESCA y CBUC, son consorcios de instituciones participantes (más o menos) en estos repositorios.

- ✓ Muchos de los servicios que ofrece/rá DuraCloud ya se están realizando, como la comprobación de la integridad de los ficheros...

- ✓ Pasar a un modelo híbrido podría ser interesante por temas de copias desgeolocalizadas, o si se concretara algún servicio de transformación de formatos. Estaremos atentos a la evolución del proyecto.

Conclusiones

- ✓ El cloud es una tendencia en auge en las TIC.
- ✓ Existen diferentes tipos de *cloud* para distintas necesidades (público, privado, híbrido, SaaS, PaaS, IaaS).
- ✓ El *data deluge* ha ayudado a acercar el cloud a los repositorios.
- ✓ DuraCloud es la solución de DURASPACE para acercar el *cloud* a DSpace y Fedora.
- ✓ Eprints dispone de *plug-ins* de almacenamiento *en cloud*.

Conclusiones

- ✓ En DuraCloud, aunque el objetivo son servicios de almacenamiento y computación en la nube, los primeros son los más avanzados con finalidades de preservación.
- ✓ Estamos aun en el principio de DuraCloud, pilotos, beta. Faltan *plug-ins* para los repositorios, SLAs, precios, casos de éxito, etc.
- ✓ Para seguir su evolución, en el próximo Open Repositorios de Madrid seguro que habrán novedades.

Agenda

- ✓ Introducción
- ✓ ¿Qué es el *cloud*?
- ✓ Tipología
  - SaaS, PaaS, IaaS
  - Público, privado, híbrido
- ✓ Participantes
- ✓ Repositorios en el *cloud*
  - Duracloud
  - EPrints storage *plug-ins*
- ✓ Conclusiones
- ✓ Referencias

Referencias

- ✓ *Above the Clouds: A Berkeley View of Cloud Computing*, Michael Armbrust et alt., UC Berkeley Reliable Adaptative Distribute Systems Laboratory, febrero 2009 (en línea en *http://eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28-pdf*).

- ✓ Information Factories, George Gilder, Wired, octubre de 2006 (en línea en http://www.wired.com/wired/archive/14.10/cloudware.html).

- ✓ Prólogo de Vinton Cerf de *Todo va a cambiar*, Enrique Dans, Ediciones Desto, 2010 (en línea en *http://filesocial.com/937y410*).

Referencias

- ✓ DuraCloud *http://www.duraspace.org/duracloud.php*

- ✓ *DuraCloud: Managing Durable Data in the Cloud.* Michele Kimpton. NDIIPP Washington, DC, junio 2009 (en línea en *http://www.duraspace.org/documents/DuraCloudNDIIPPJune09.ppt*).

- ✓ *DuraCloud Frequently asked questions*. Michele Kimpton y Bill Branan, octubre 2009 (en línea en *http://www.fedora-commons.org/confluence/display/duracloudpilot/Frequently+asked+questions*).

Referencias

- ✓ *Repositories and the Cloud*. 23 febrero de 2010 en Londres (presentaciones y videos en línea en *http://userv.org.uk/events/repcloud*).

  - *Duracloud – Open technologies and services for managing durable data in the cloud*, Michele Kimpton, DuraSpace.
  - *Cloud Services for Repositories*, Alex Wade, Microsoft.
  - *Eprints and the Cloud*, Les Carr, University of Southampton.
  - *Cloud based Projects at Belfast e-Science Centre*, Terry Harmer, Belfast e-Science Centre.

Más referencias...

http://www.youtube.com/watch?v=QJncFirhjPg

Cloud Computing Explained

Confused about the term "**Cloud Computing**"? Want to be "with the times" when you talk about new technology buzzwords? This video boils down a ...

★★★★★ 1 year ago 143,949 views HighT3chDad

http://www.youtube.com/watch?v=XdBd14rjcs0

Cloud Computing Plain and Simple

** Not affiliated with Common Craft. If you're interested in Common Craft, visit: www.commoncraft.com ** rPath takes the confusion out of **cloud ...**

★★★★☆ 1 year ago 115,885 views rhirschfield

http://www.youtube.com/watch?v=n9LmzsaO698

Cloud Computing HD

Short video showing how **cloud computing** will help us in the future.

★★★★½ 2 months ago 3,465 views GoogleChannelUK

Más referencias...

http://www.youtube.com/watch?v=Cl6XFZH5aWU

Richard **Stallman** and Marcelo d'Elia Branco on Free Software,

Richard **Stallman** (GNU Founder) and Marcelo d'Elia Branco (Free Software leader in Brasil) talk about Software as a Service, **Cloud Computing** and **...**

★★★★½ 8 months ago 1,486 views pau8000

http://www.youtube.com/watch?v=VjfaCoA2sQk

Hitler and **Cloud Computing** Security

Hitler learns a painful lesson about **Cloud Computing** security.

★★★★★ 3 days ago 20,089 views gwcstudio

http://www.youtube.com/watch?v=TcTnGAQJ7gE

Rackspace **Public Cloud** vs. **Private Cloud** Computing

The difference between **private** and **public** storage **clouds** is simple. Where is the **cloud** deployed? A **public cloud** is offered as a service, usually **...**

no rating 1 week ago 10 views TheRackspaceCloud

CENTRE DE SUPERCOMPUTACIÓ
DE CATALUNYA

CESCA

ANELLA CIENTÍFICA · CATNIX · TDX · RACO · RECERCAT · mdx · JOCS · TAPC · TSIUC · TERAFLOP