

Procesamiento Masivo de datos

Alex Di Genova

24/08/2022

Outline

- Big Data
- Sistemas distribuidos
 - Tipos y arquitecturas.

Big Data

Visión global



Byte B	10^0	1
Kilobyte	$KB10^3$	1,000
Megabyte	$MB10^6$	1,000,000
Gigabyte	$GB10^9$	1,000,000,000
Terabyte	$TB10^{12}$	1,000,000,000,000
Petabyte	$PB10^{15}$	1,000,000,000,000,000
Exabyte	$EB10^{18}$	1,000,000,000,000,000,000

- Astronomia
- Genomica
- Redes Sociales
- Youtube



Stephens, Zachary D., et al. "Big data: astronomical or genomic?." *PLoS biology* 13.7 (2015): e1002195.



Visión global

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction Real-time processing Massive volumes	Topic and sentiment mining Metadata analysis	Limited requirements	Heterogeneous data and analysis Variant calling, ~2 trillion central processing unit (CPU) hours All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

- Astronomia
- Genomica
- Redes Sociales
- Youtube



Stephens, Zachary D., et al. "Big data: astronomical or genomic?." *PLoS biology* 13.7 (2015): e1002195.

Visión global



Octubre 2020,
Sebastian
Steudtner
(German)
26 metros

- Astronomia
- Genomica
- Redes Sociales
- Youtube
- Infraestructura (computadores, redes)
- algoritmos (software)

Sistemas Distribuidos

Sistemas distribuidos



Gran cantidad de computadores conectados por una red de alta velocidad.

- Un sistema distribuido es una **colección de computadoras** independientes que aparece ante sus usuarios como un **solo sistema coherente**.

Sistemas distribuidos



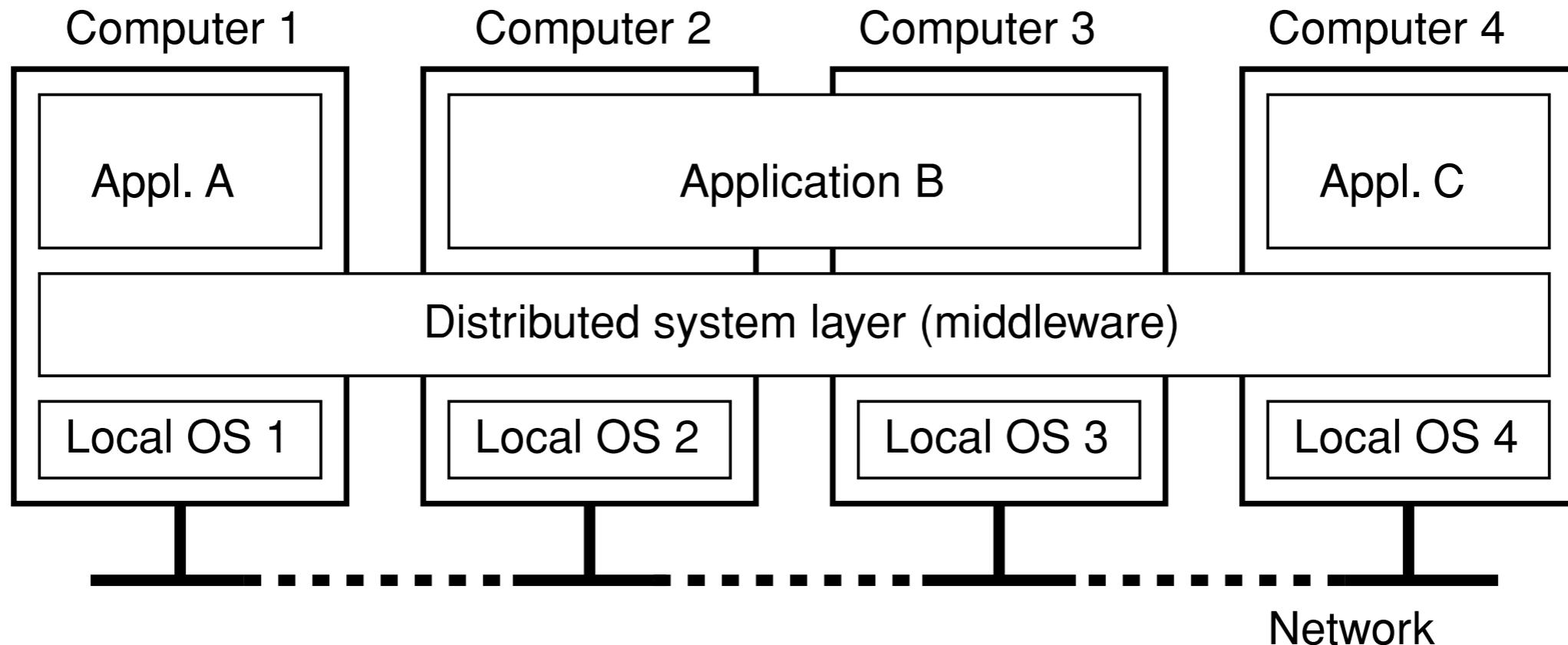
- Un sistema distribuido es una **colección de computadoras** independientes que aparece ante sus usuarios como un **solo sistema coherente**.
 - -> las **computadoras necesitan colaborar**.
 - **Cómo establecer esta colaboración** se encuentra en el corazón del desarrollo de sistemas distribuidos.
- Dentro de un sistema distribuido, podrían existir computadoras de alto rendimiento hasta pequeños nodos (**heterogéneo**).
- No se hacen suposiciones sobre la forma en que se **interconectan las computadoras**.

Sistemas distribuidos



- Un sistema distribuido es una **colección de computadoras** independientes que aparece ante sus usuarios como un **solo sistema coherente**.
- **Características**
 - Las diferencias entre las distintas computadoras y las formas en que se comunican están en su mayoría ocultas a los usuarios.
 - Los usuarios y las aplicaciones pueden interactuar con un sistema distribuido de manera consistente y uniforme
 - Relativamente fácil de expandir o escalar.
 - Alta disponibilidad.

Sistemas distribuidos

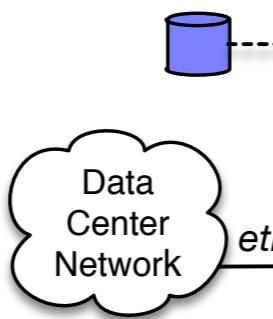


- El middleware (software) se extiende sobre varias máquinas y ofrece a cada aplicación/usuario la misma interfaz para interaccionar con el SD.
- Un sistema distribuido debe hacer que los recursos sean fácilmente accesibles (1); debería ocultar razonablemente el hecho de que los recursos se distribuyen a través de una red(2); debe ser abierto – Interface definition language (3); y debe ser escalable(4).

Sistemas distribuidos

Tipos

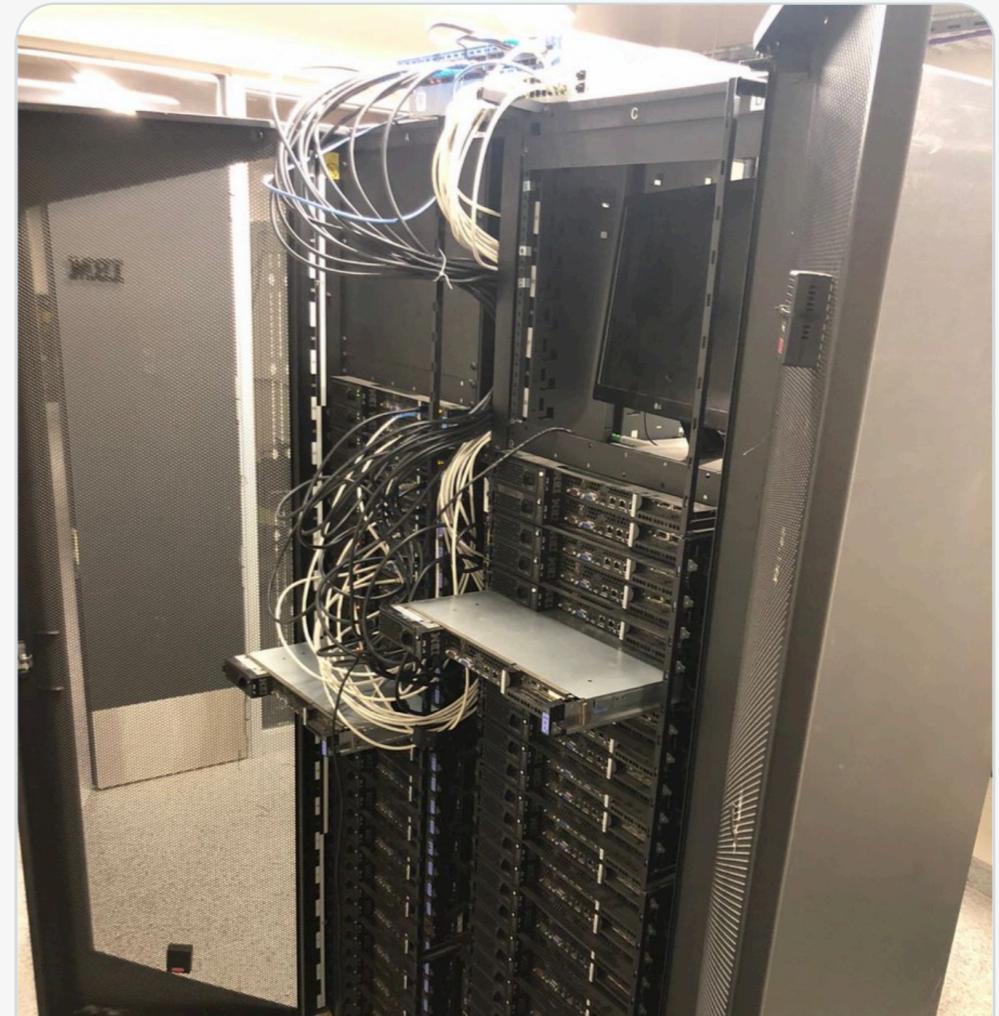
- Sistemas de computo distribuido
 - Cluster de computo
 - Hardware similar
 - Red local de alta velocidad
 - Mismo sistema operativo



⤓ You Retweeted

Raúl Valenzuela @raulrainfall · Jul 27

En @uo Higgins estamos construyendo un computador HPC para crear capacidad de cómputo en la Región de O'Higgins y abordar problemas científicos complejos en áreas como genómica y cambio climático. Proyecto liderado por @digenoma con la ayuda de @CMMUChile y @NLHPC



Tod@s invitad@s a contribuir a configurar el primer cluster de la UOH y la Región de O'Higgins

The screenshot shows the GitHub organization page for 'HPC-UOH'. The main header includes the GitHub logo, user profile picture, and a search bar. Below the header, the organization's logo (a stylized 'UOH'), name, location (Chile), website (http://www.uoh.cl), and email (hpc@uoh.cl) are displayed. A 'Follow' button is also present.

Overview tab is selected. The README file contains a short description of the HPC group, mentioning they are Chilean-based and part of the DTI department. It also lists hardware specifications: 66 nodes, 528 cores, 3.1Tb RAM (48Gb/node), Infiniband Connection (40Gb/s), and 192 Tb of storage. A large image of two server racks is shown.

The 'People' section shows three profile icons and a 'Invite someone' button.

The 'Repositories' section lists two public repositories:

- .github** [Public] - Description of our HPC-UOH group. Last updated 2 minutes ago.
- installation_configuration** [Private] - This repository contains useful documentation about the first configuration and installation process of the UOH cluster. Last updated on Mar 30.

At the bottom, a horizontal bar shows other open files: 'Architecting Mod....pdf', 'Distributed syste....pdf', 'Formato_antec....docx', and 'Formato_presup....xlsx'.

HPC-UOH

Installation & configuration

This repository contains useful documentation about the first configuration and installation process of the UOH cluster.

Work plan 18/03/2022

- ¿Chequeo detallado de maquinas? (encendido, discos, RAM ..) [UOH-DTI] (11H ~ 2D)
- Instalación de equipos y conectividad en el rack IBM. [UOH-DTI] (1D)
- Configuración de redes: firewalls, switches. [UOH-DTI](#)
- Instalación y configuración DNS. [UOH-DTI] (1D)
- Distribución de Linux a utilizar, Ubuntu, ¿Centos? PXE boot instalación?(Depende de la controladora IB OFED LTS) [UOH-DTI] (1D)
- Instalación y configuración LDAP (Manejo centralizado de usuarios). [UOH-DTI] (1D)
- Habilitar acceso remoto al clúster mediante VPN. [UOH-DTI](#)
- Instalación y configuración de Sistema de Virtualización (¿opcional?).
- Sistema de despliegue de imágenes. xCAT? [UOH-NLHPC]
- Instalación y configuración de EasyBuild (<https://easybuild.io/>). [UOH-NLHPC]
- Instalación y configuración de Ganglia. [UOH-NLHPC]
- ¿Instalación y configuración de FS. BeeGFS? Lustre? [UOH-NLHPC]
- Instalación y configuración de sistema gestor de recursos. SLURM? [UOH-NLHPC]
- ¿La configuración podrá ser en remoto con apoyo local? [UOH-NLHPC]

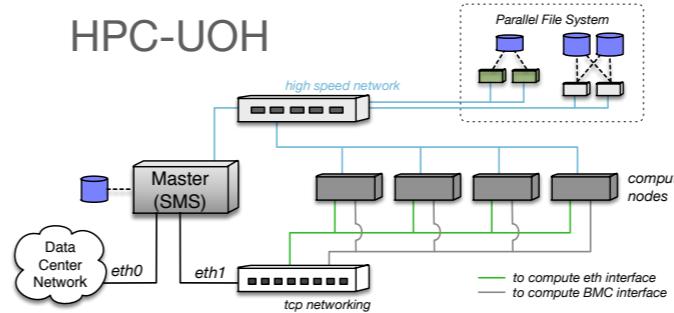
Linux y ganas de aprender!!!

Sistemas distribuidos

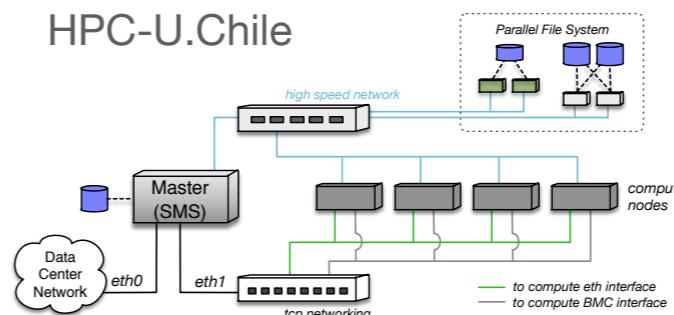
Tipos

- Sistemas de computo distribuido
 - Cluster de computo
 - Hardware similar
 - Red local de alta velocidad
 - Mismo sistema operativo
 - Grid de computo
 - Alto grado de heterogeneidad
 - Federación de sistemas de computo (cluster)

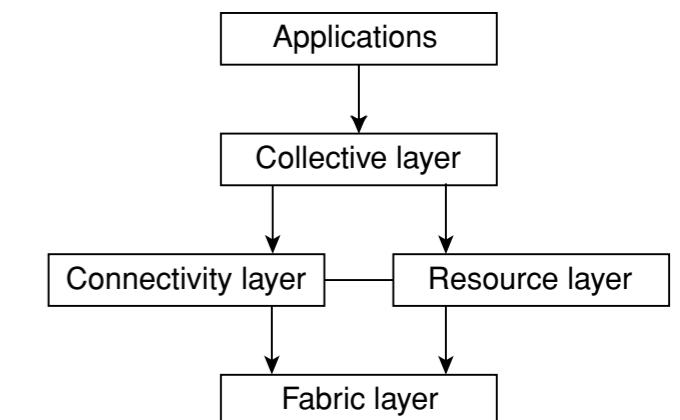
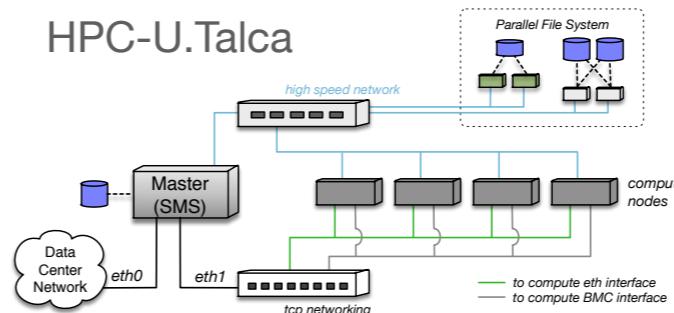
Universidades (organización virtual)



HPC-U.Chile



HPC-U.Talca



- Fabric layer:
 - proporciona interfaces a los recursos locales en un sitio específico
- Connectivity layer
 - protocolos de comunicación para soportar transacciones de red que abarcan el uso de múltiples recursos.
- Resource layer
 - responsable de administrar un solo recurso
- Collective layer:
 - maneja el acceso a múltiples recursos y generalmente consiste en servicios para el descubrimiento de recursos, asignación y programación de tareas

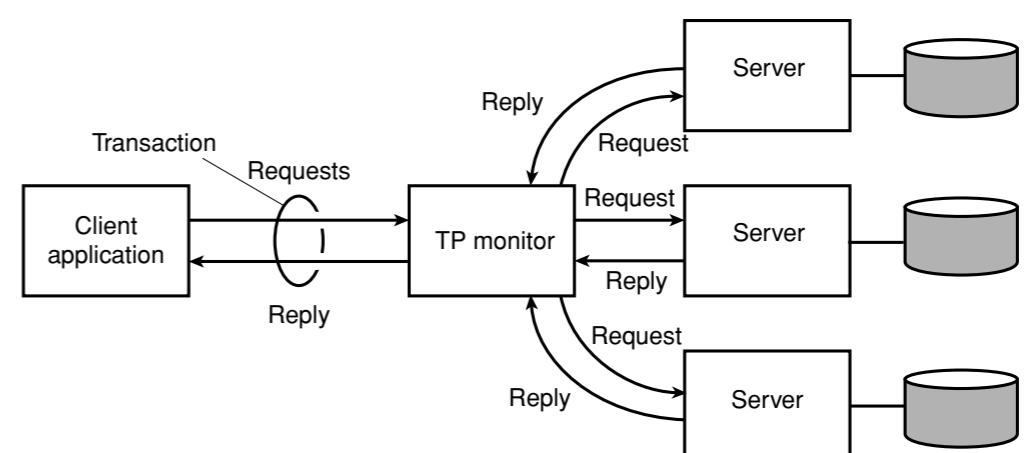
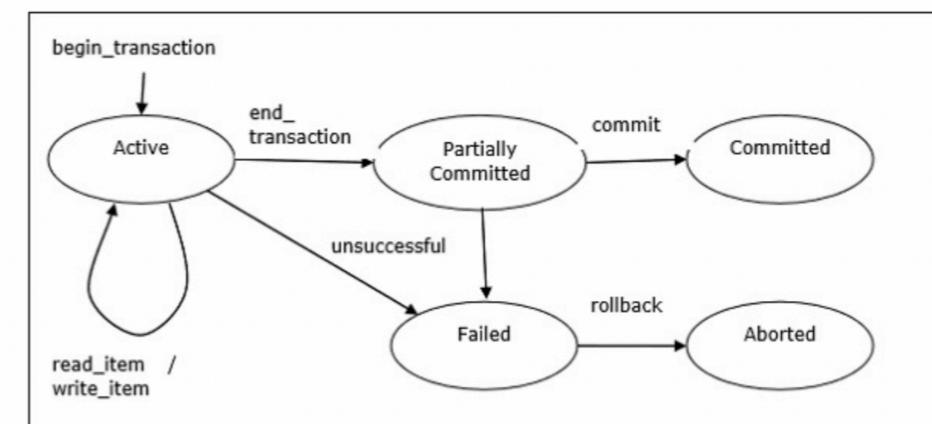
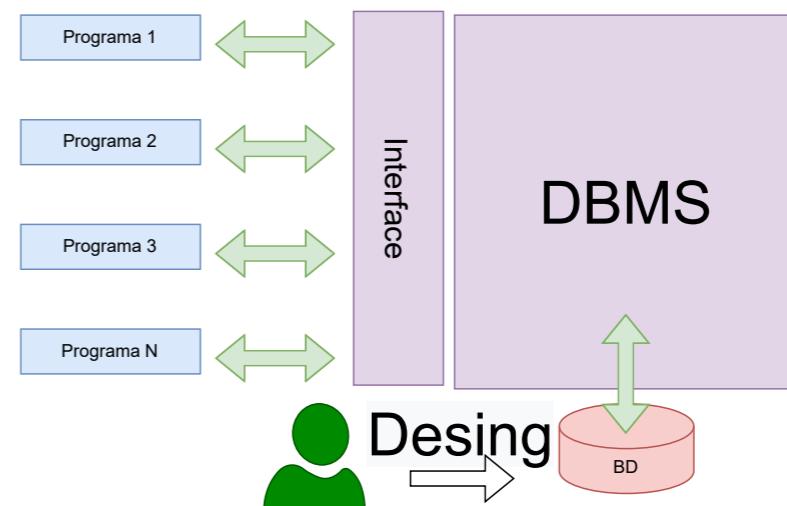


HTCondor
Software Suite

Sistemas distribuidos

Tipos

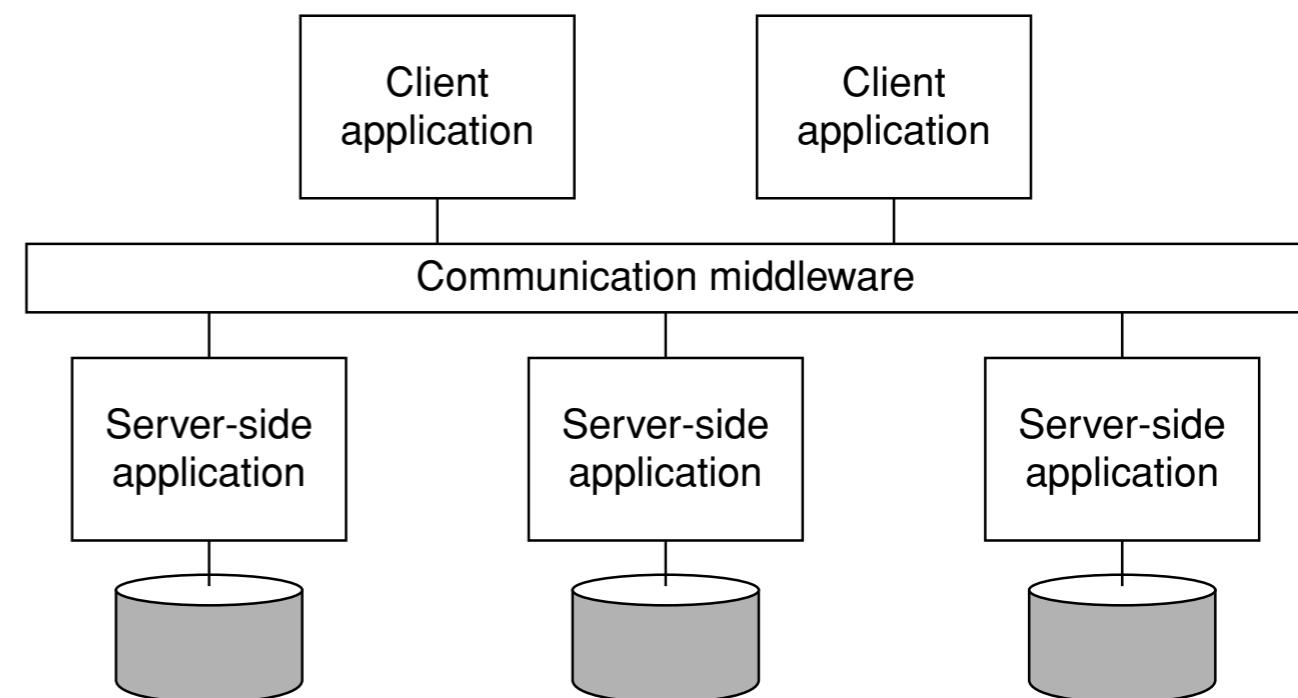
- Sistemas de información distribuido
 - un servidor que ejecuta una aplicación (que a menudo incluye una base de datos) y la pone a disposición de programas remotos, llamados clientes.
 - Sistemas de procesamiento de transacciones
 - La idea clave es que se ejecutarán todas o ninguna de las solicitudes.
 - Propiedades ACID (Atomic, Consistent, Isolated, Durable)
 - monitor de procesamiento de transacciones
 - Permite que una aplicación acceda a múltiples servidores/bases de datos mediante un modelo de programación transaccional.



Sistemas distribuidos

Tipos

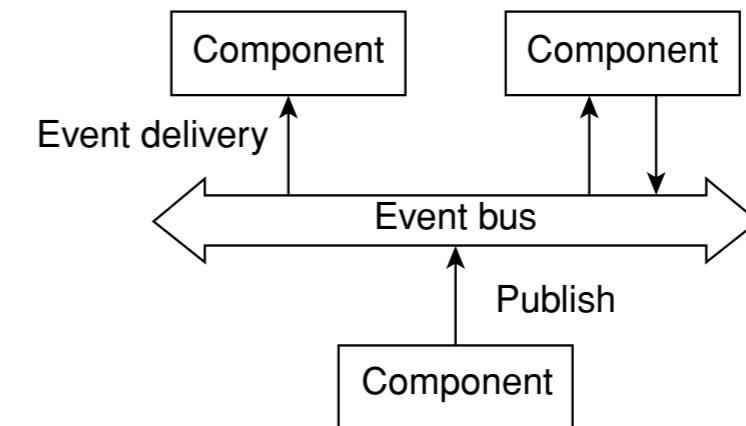
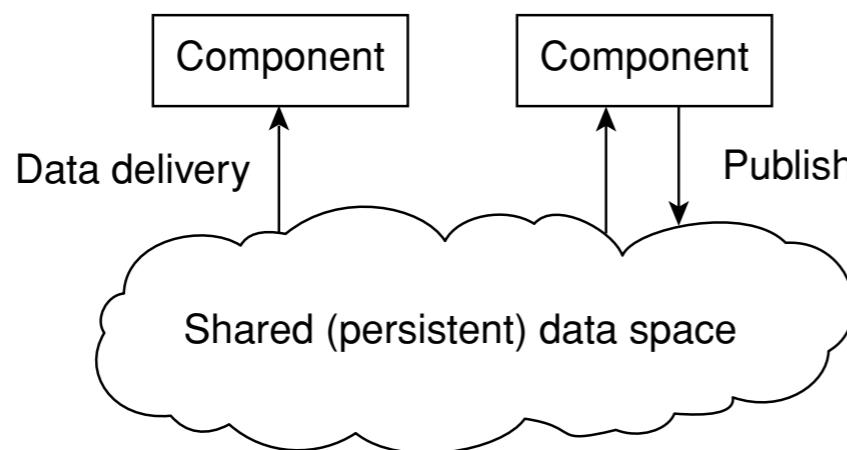
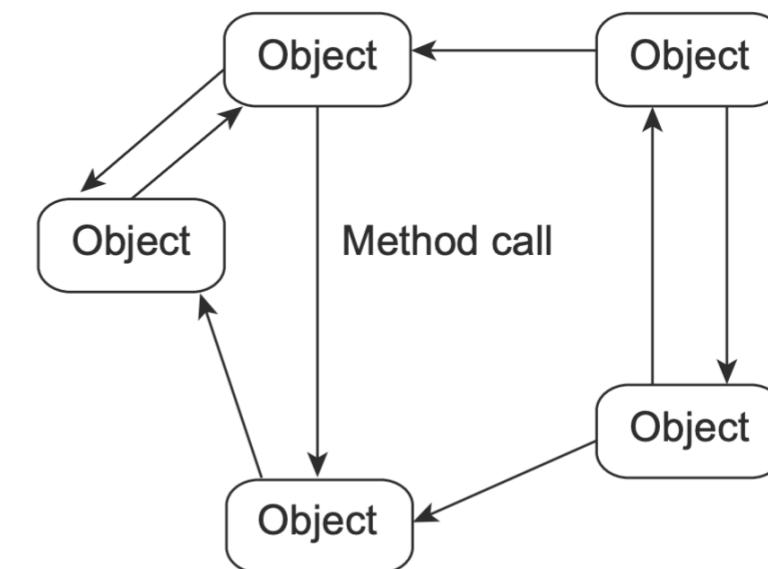
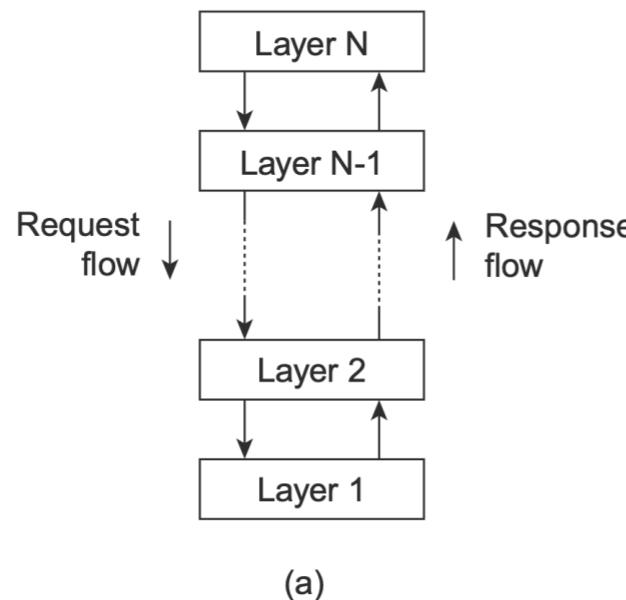
- Sistemas de información distribuido
 - Sistemas de procesamiento de transacciones
 - Sistemas de integración de aplicaciones.
 - Comunicación entre aplicaciones.
 - Idea: aplicaciones intercambian directamente información.
 - Invocaciones remotas de métodos (RMI)



Sistemas distribuidos

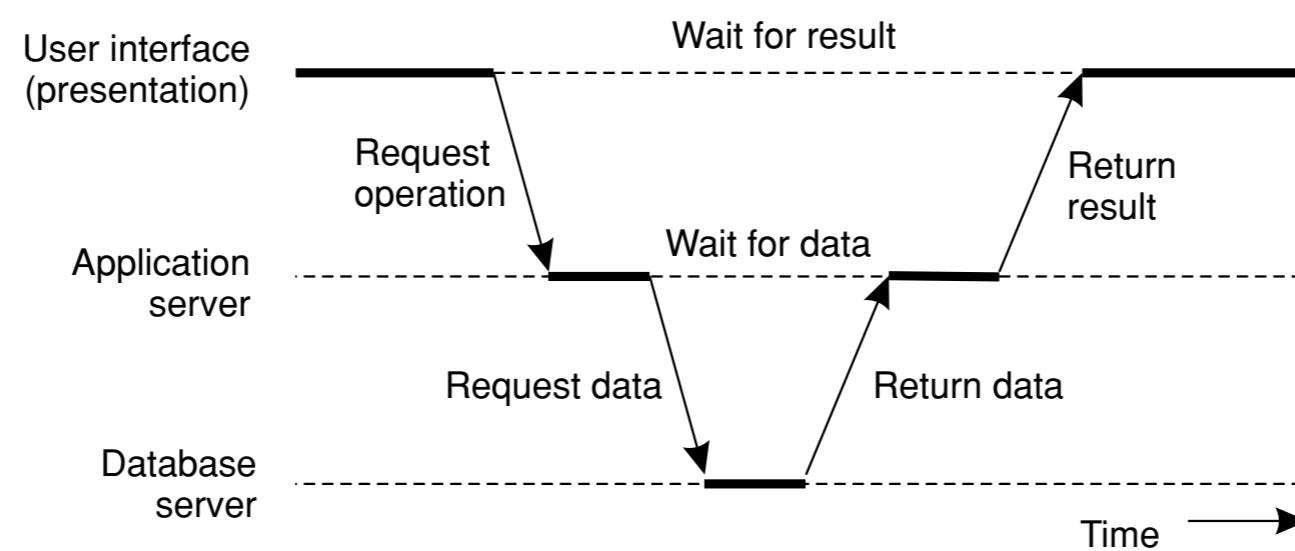
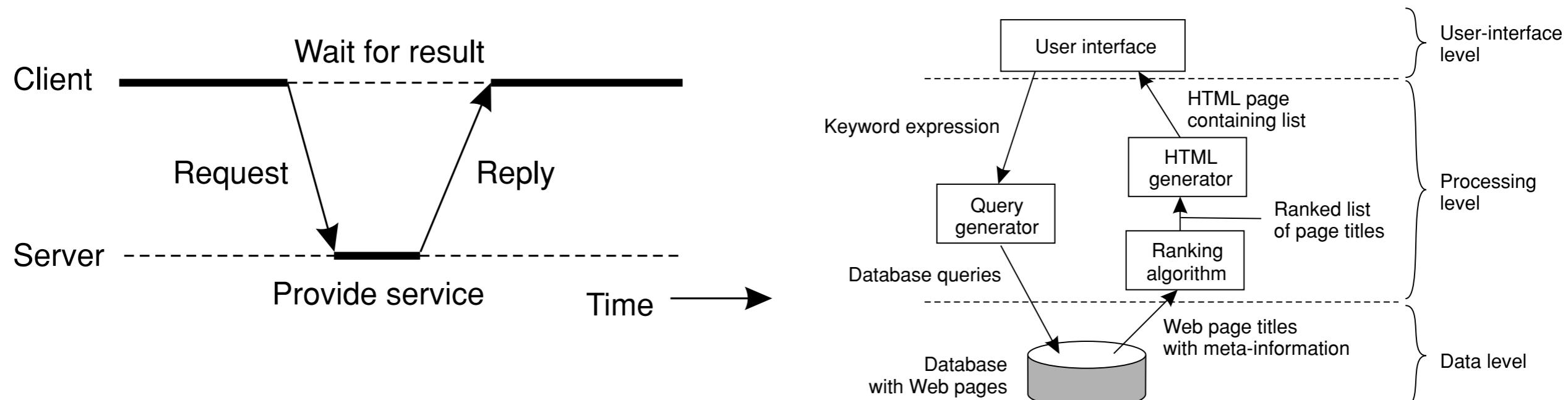
Arquitecturas

- Necesitamos definir cómo deben organizarse los diversos componentes de software y cómo deben interactuar.



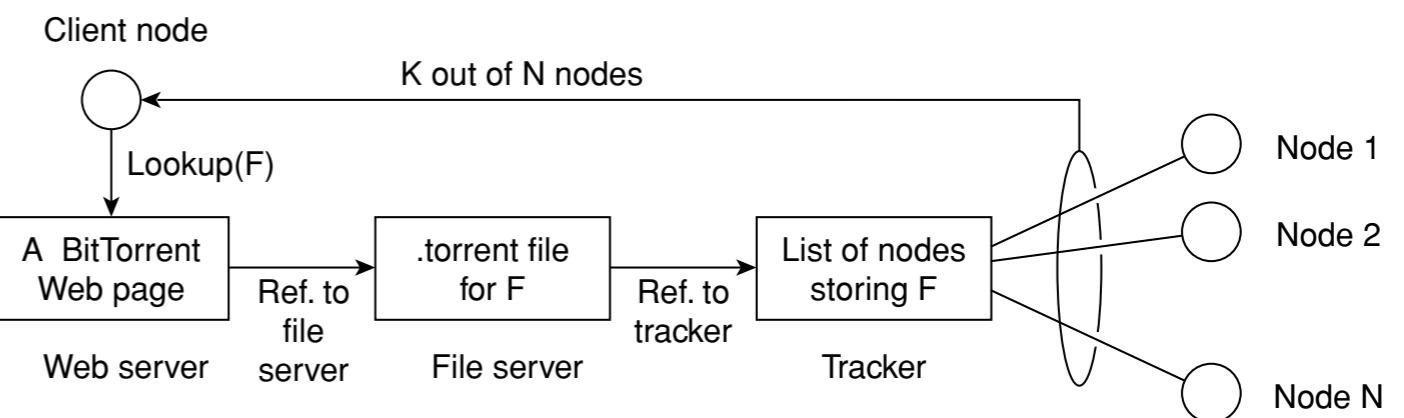
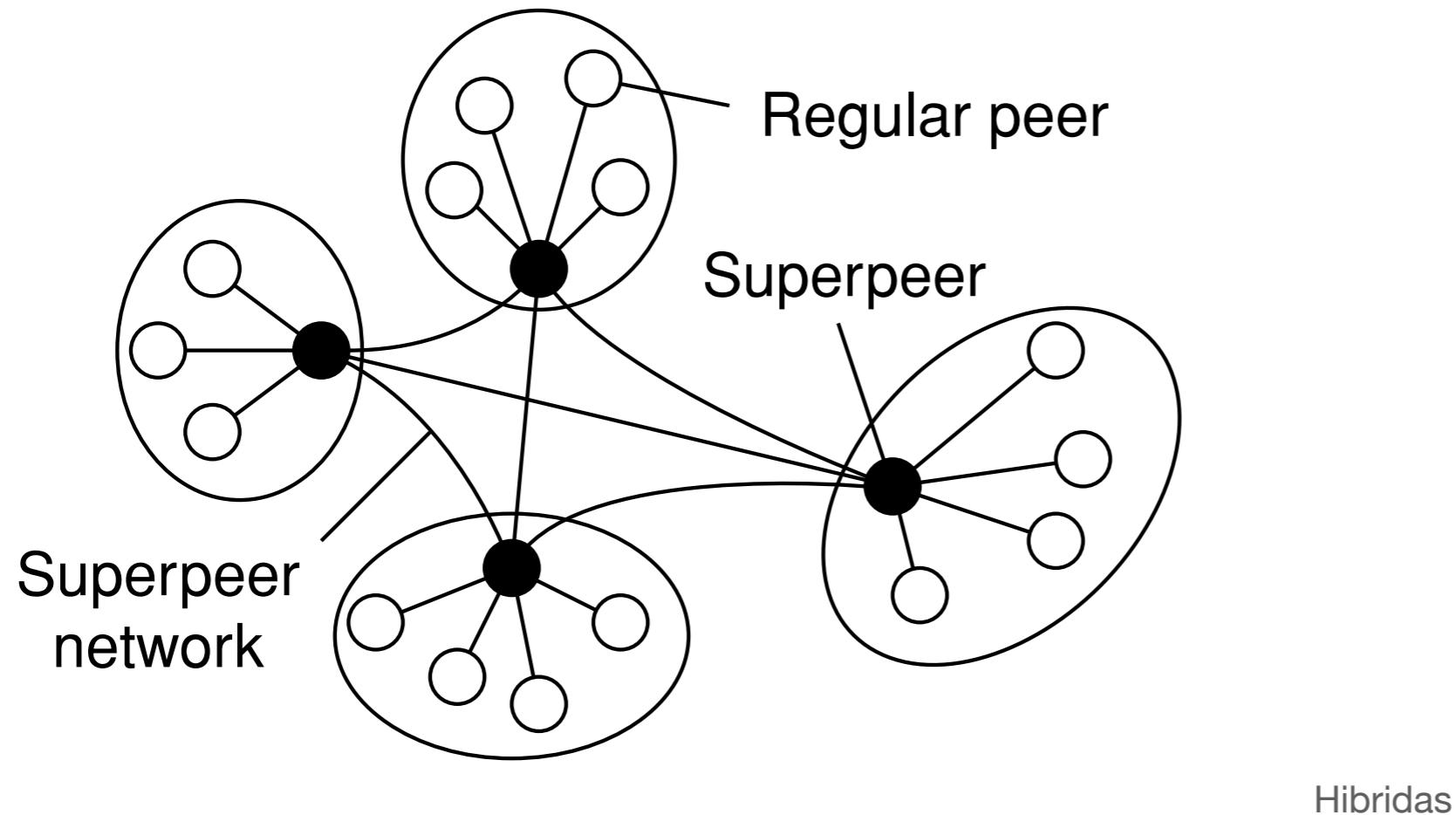
Sistemas distribuidos

Arquitecturas centralizadas



Sistemas distribuidos

Arquitecturas no-centralizadas



Visita a Cluster UOH



Consultas?

Consultas o comentarios?

Muchas gracias