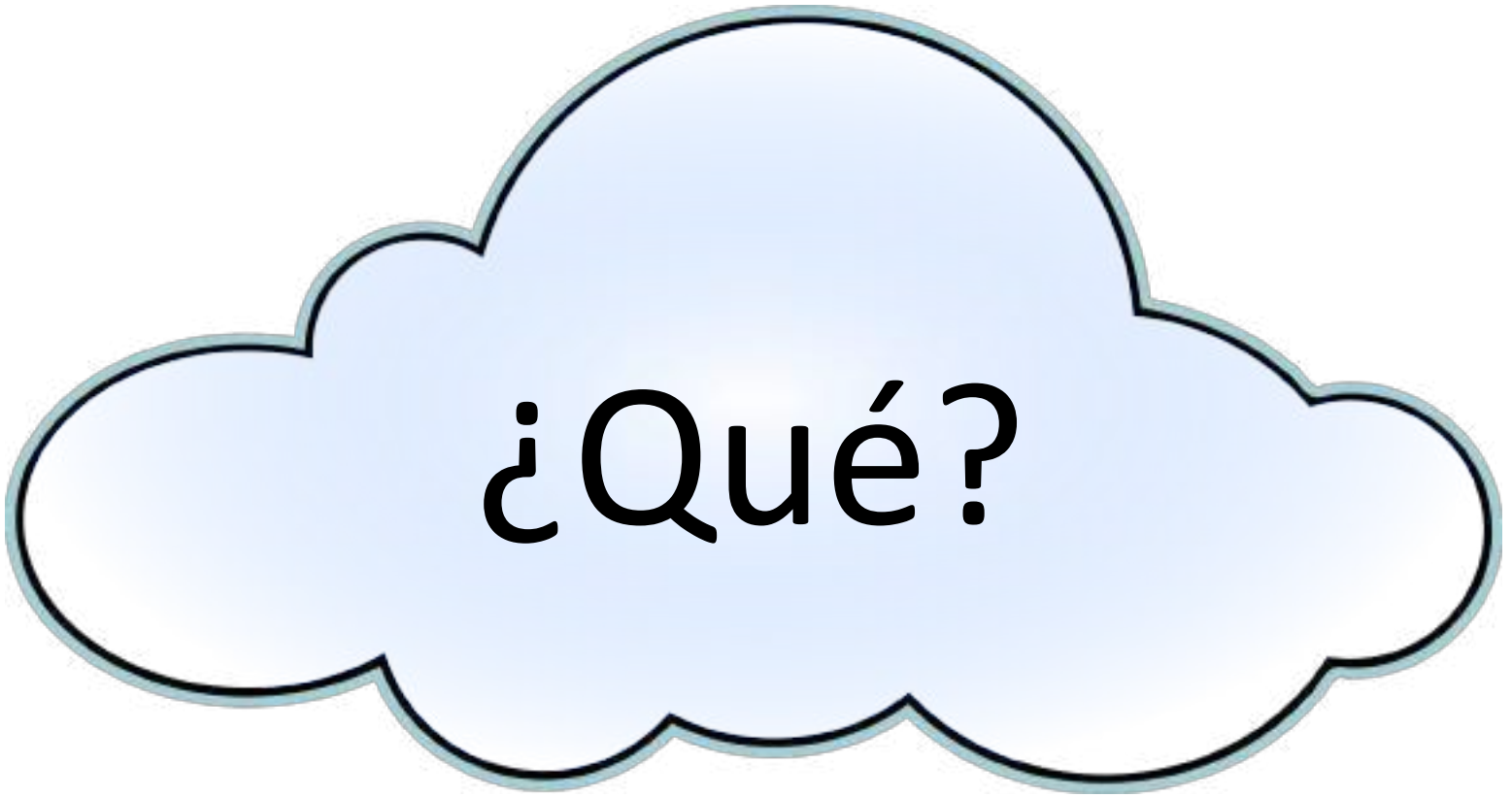
La Nube se hace Niebla

Juan Carlos López


¿Qué?

Almacenamiento

Acceso a servicios software

Plataforma de desarrollo

Infraestructura

(almacenamiento y cómputo)

Reducción de costes (uso y gestión)

Colaboración

Competitividad

Flexibilidad

Eficiencia

Calidad

Flexibilidad

Seguridad (recuperación de desastres)

Acceso

Recursos centralizados y remotos

Necesidad de Internet (de calidad)

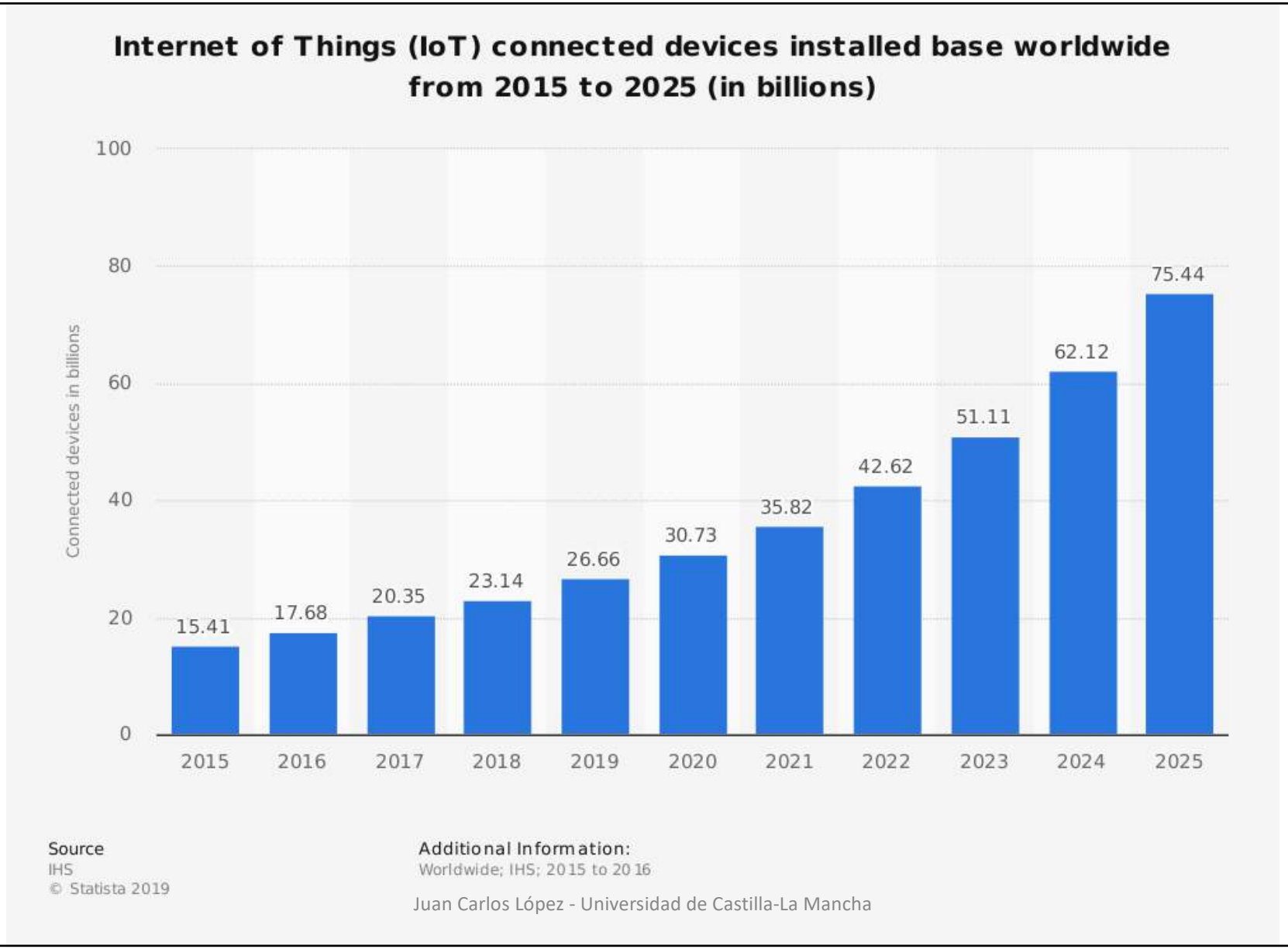
Fiabilidad - Seguridad – Privacidad

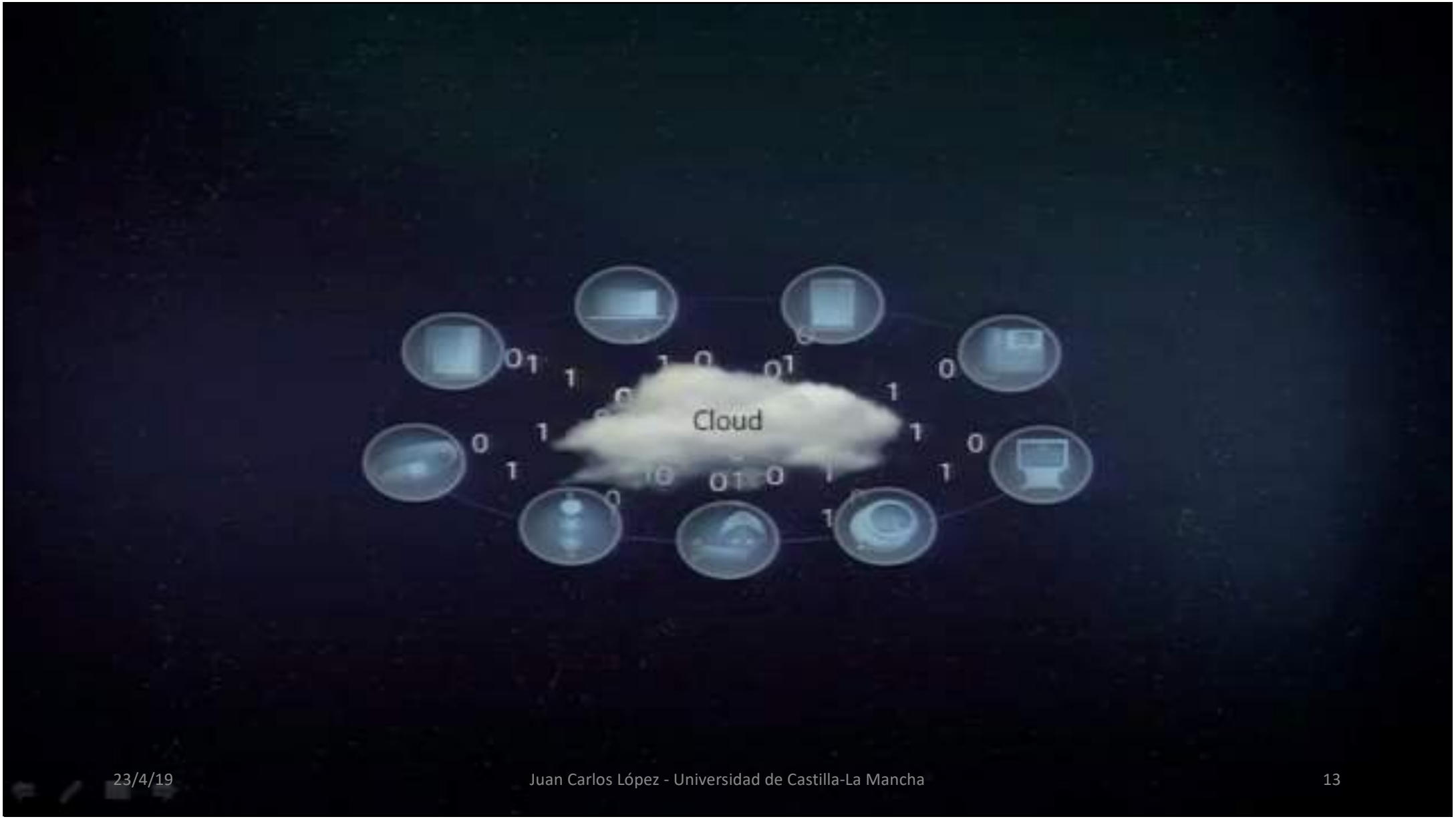
Escalabilidad a largo plazo

Conectividad


¿Qué pasa durante un minuto en internet?
facebook
NETFLIX
WhatsApp
701,389
Inicios de sesión
69,444
horas de reproducción
150 millones
de correos electrónicos enviados
20.8+
millones de mensajes enviados
YouTube
2.78
millones de reproducciones
1,389
viajes
UBER
tinder
972,222
swipes
527,760
fotos compartidas
2.4
millones de búsquedas
45
15
30
Google
51,000
descargas
Download on the App Store
38,052
horas de música
$203,596
dólares en ventas
amazon
Spotify
120+
nuevas cuentas
1.04
millones
de loops
347,222
tweets
LinkedIn
Vine
38,194
posteos
Instagram
FUENTE:
EXCELACOM
©2016 Excelacom, Inc.
paréntesis.com

Internet of (smart) Things


Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)
Connected devices in billions
100
80
60
40
20
0
15.41
17.68
20.35
23.14
26.66
30.73
35.82
42.62
51.11
62.12
75.44
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Source
IHS
© Statista 2019
Additional Information:
Worldwide; IHS; 2015 to 2016


Cloud


¿Dónde?

Google Cloud Regions and Zones

https://cloud.google.com/about/locations/


Paris City Guide
Restaurants
Hotels
Entertainment
Sightseeing
ATMs

Consumo

computing

INICIO SEGURIDAD ANALYTICS CLOUD MERCADO TI MOVILIDAD INFRAESTRUCTURAS

INFORMES DE VALOR TIC

DESCARGA GRATUITA

INFORMES > INFRAESTRUCTURAS

Los CPD de todo el mundo consumen 30.000 millo vatios, un 1,5% de todo el consumo mundial

Ser "verde" va más allá de un mero mensaje de marketing, la reducción de costes es ahora u lleva implícito renovar servidores, almacenamiento y conectividad.

ecoavant.com

La actualidad del medio ambiente

16 de mayo

Buscar

PORTADA | ACTUALIDAD | MEDIO AMBIENTE | NATURALEZA | CONSUMO | SOSTENIBILIDAD | CONTAN

Ud. está en: Portada | Consumo | La nube es tóxica

COMPARTE > Facebook 46 Twitter WhatsApp Meneame

› Consumo Jueves, 07 de febrero de 2013 Pau Ruiz

La nube es tóxica

Los servidores que almacenan los datos de Internet consumen y desperdician grandes cantidades de energía por la acumulación de información, el temor a un corte del servicio y el sobrecalentamiento

≡ EL PAÍS

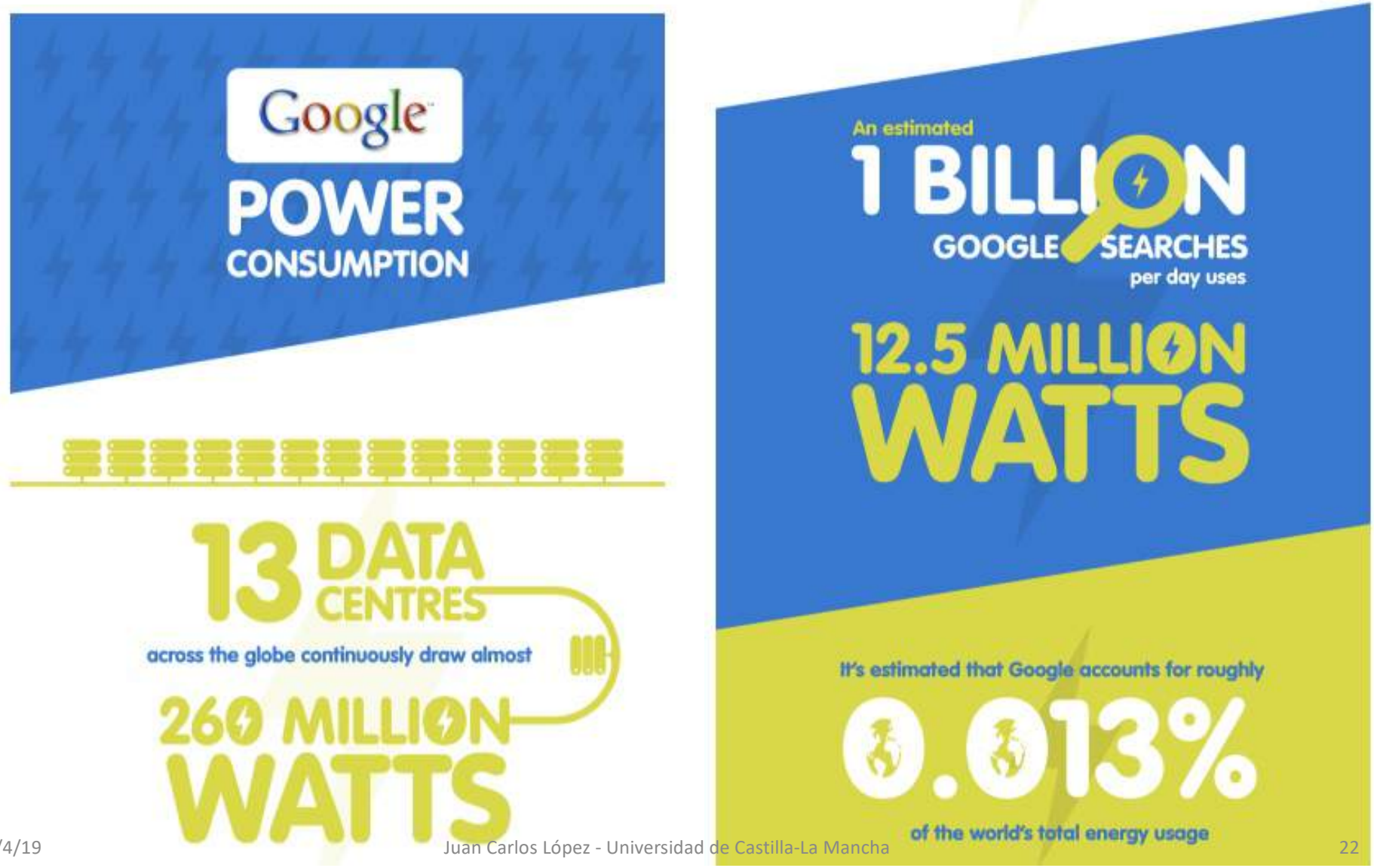
TECNOLOGÍA

MÓVILES REDES SOCIALES BANCO DE PRUEBAS RETINA MERISTATION

¿Y dónde vamos a enchufar el 'Internet de las cosas'?

El mundo de los objetos conectados está a la vuelta de la esquina. Pero esta nueva tecnología se enfrenta a un desafío clave: producir la energía necesaria para ser sostenible

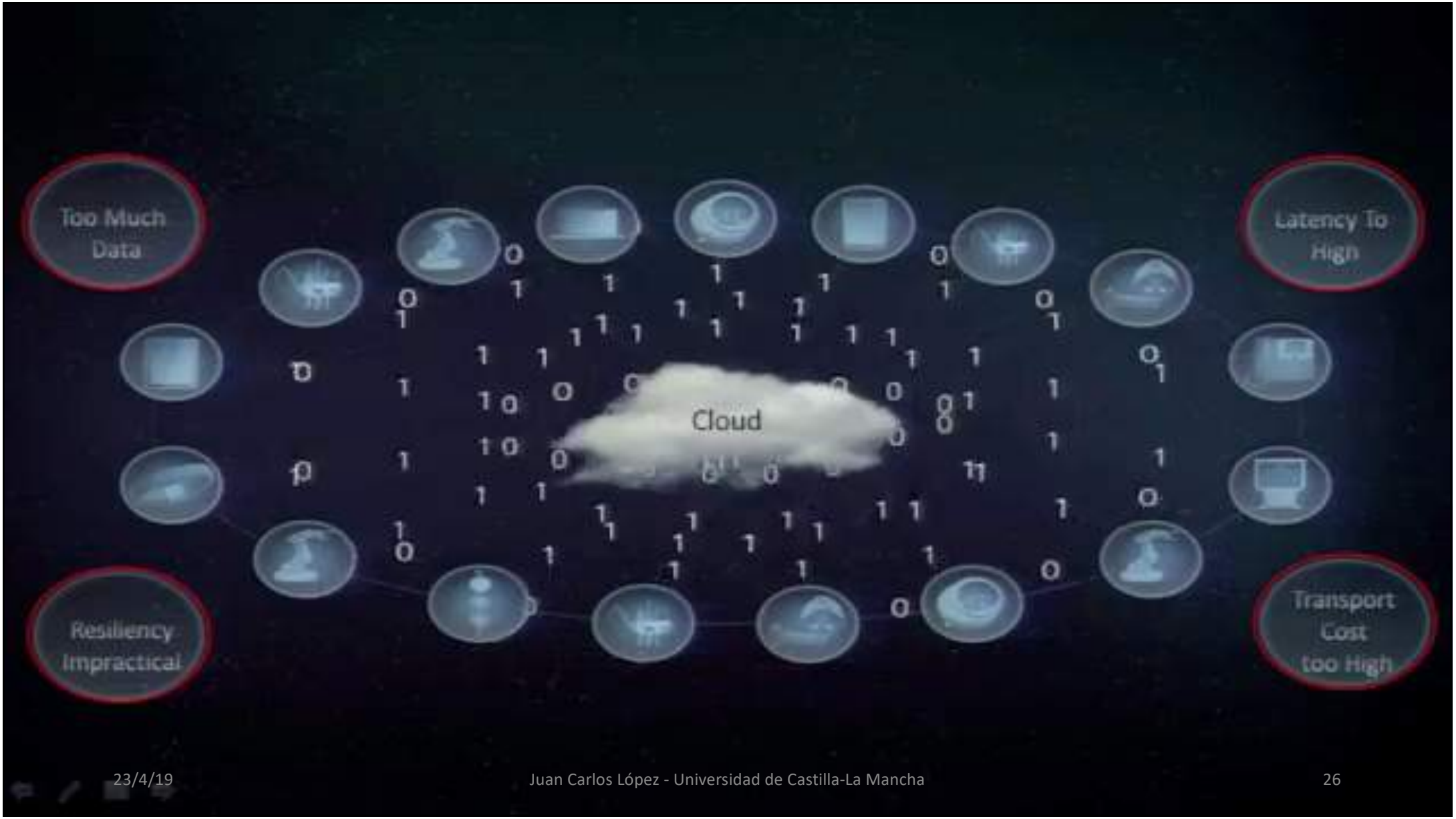

¿Cómo?

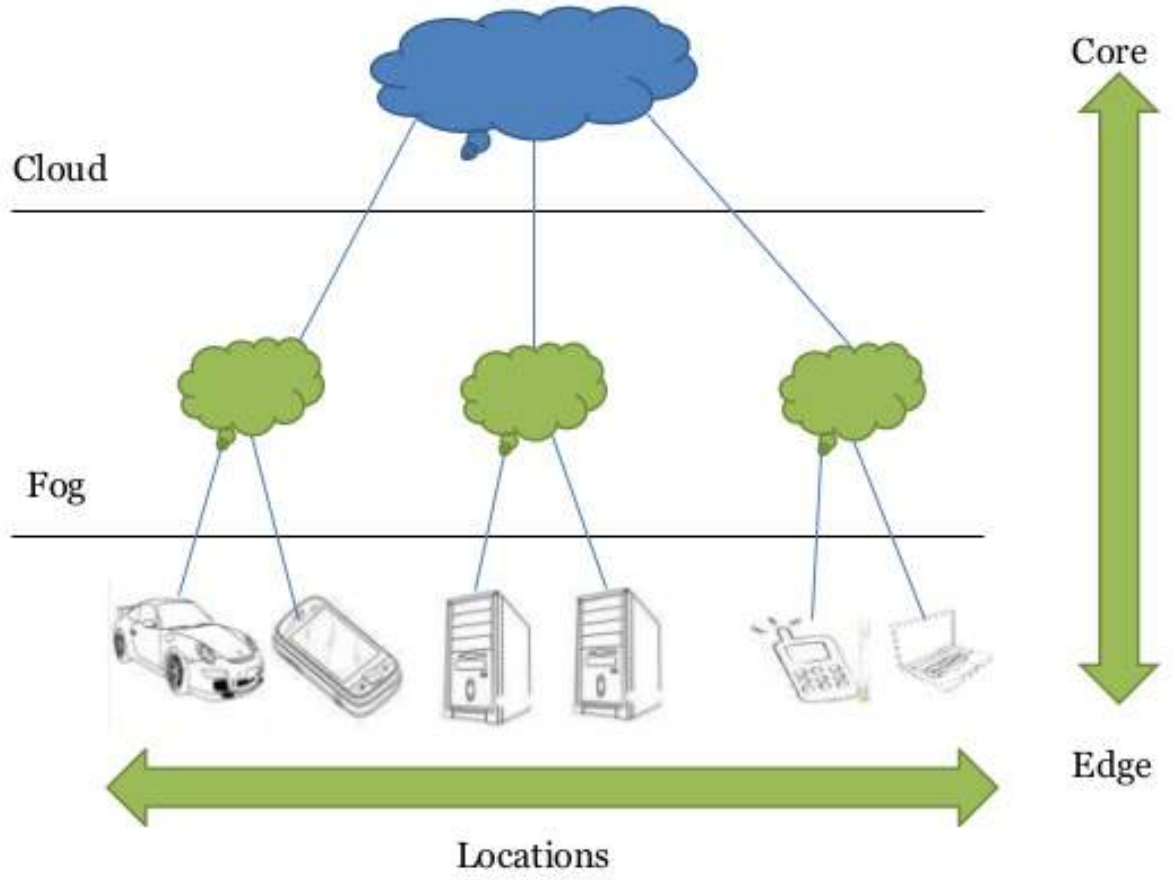

Google
POWER
CONSUMPTION
13 DATA CENTRES
across the globe continuously draw almost
260 MILLION WATTS
An estimated
1 BILLION
GOOGLE SEARCHES
per day uses
12.5 MILLION WATTS
It's estimated that Google accounts for roughly
0.013%
of the world's total energy usage

La nube está en la Tierra...

…y es el sexto “país” del mundo en gasto eléctrico

Niebla


Too Much Data
Latency To High
Cloud
Resiliency Impractical
Transport Cost too High


Core
Cloud
Fog
Edge
Locations

En la era de lo *Smart*…

…be intelligent

¿Es posible, gracias al avance computacional (prestaciones y consumo) de los dispositivos finales…

…inferir localmente?

Modelo actual

Todo el proceso de inferencia se realiza en la nube

Oye, Siri
Ok, Google

- Grabar vídeo
- Comprimir vídeo
- Enviar audio a los servidores de google en … (?)
- Recibir los resultados de la consulta

Modelo propuesto

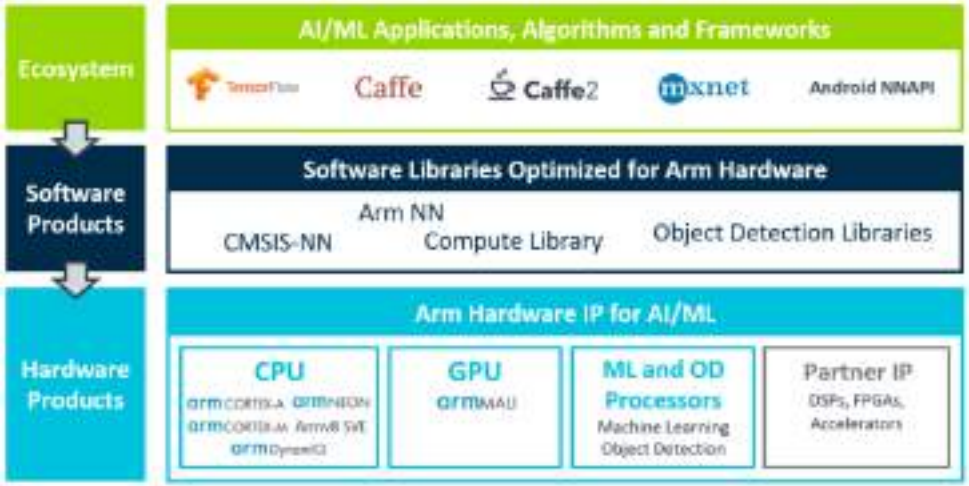
Realizar la **inferencia** en el dispositivo final y usar la nube en caso de requerir mayor precisión

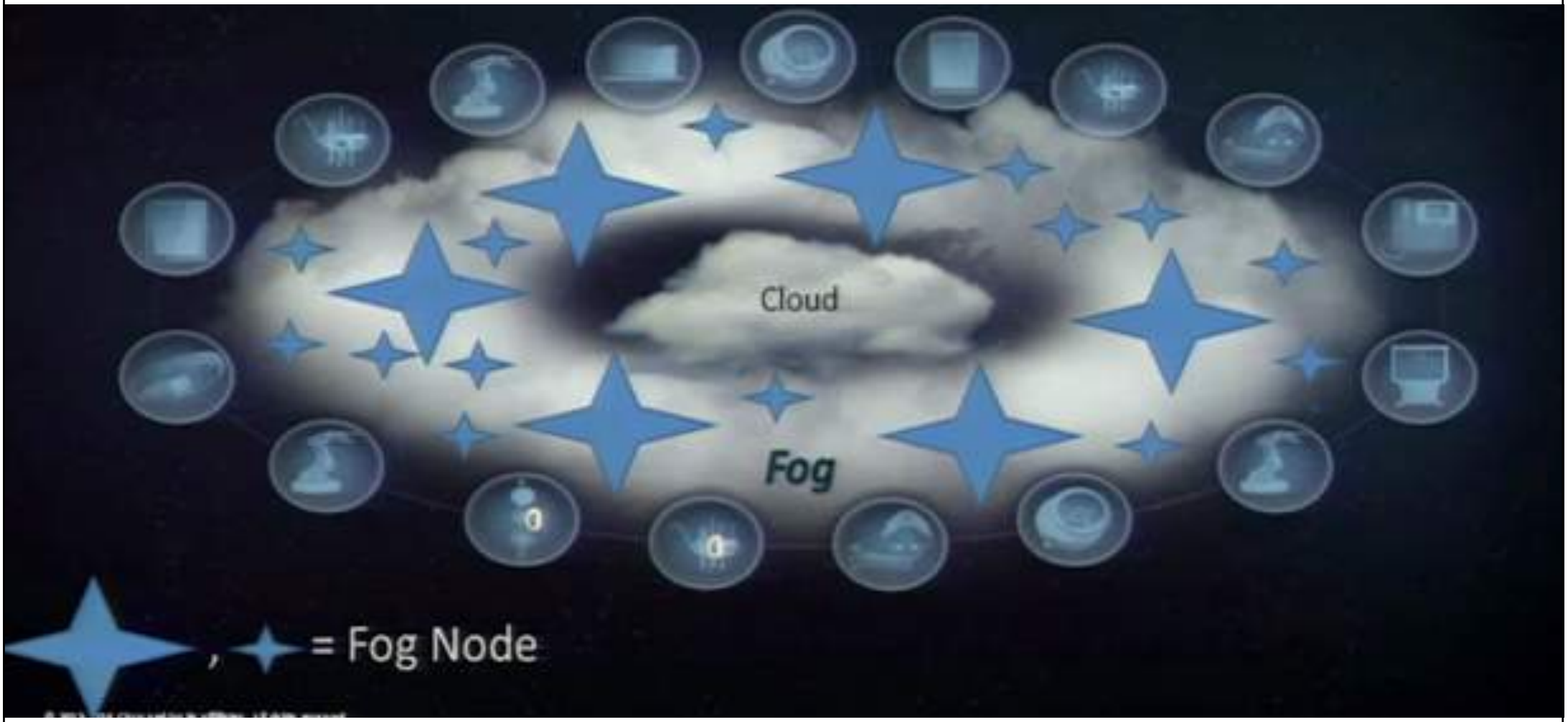
Oye, Siri
Ok, Google

- Grabar vídeo
- Comprimir vídeo
- Utilizar modelo pre-entrenado
- Evaluar el resultado
- En caso de necesitar más precisión derivar a la nube

Mejorando la inferencia local

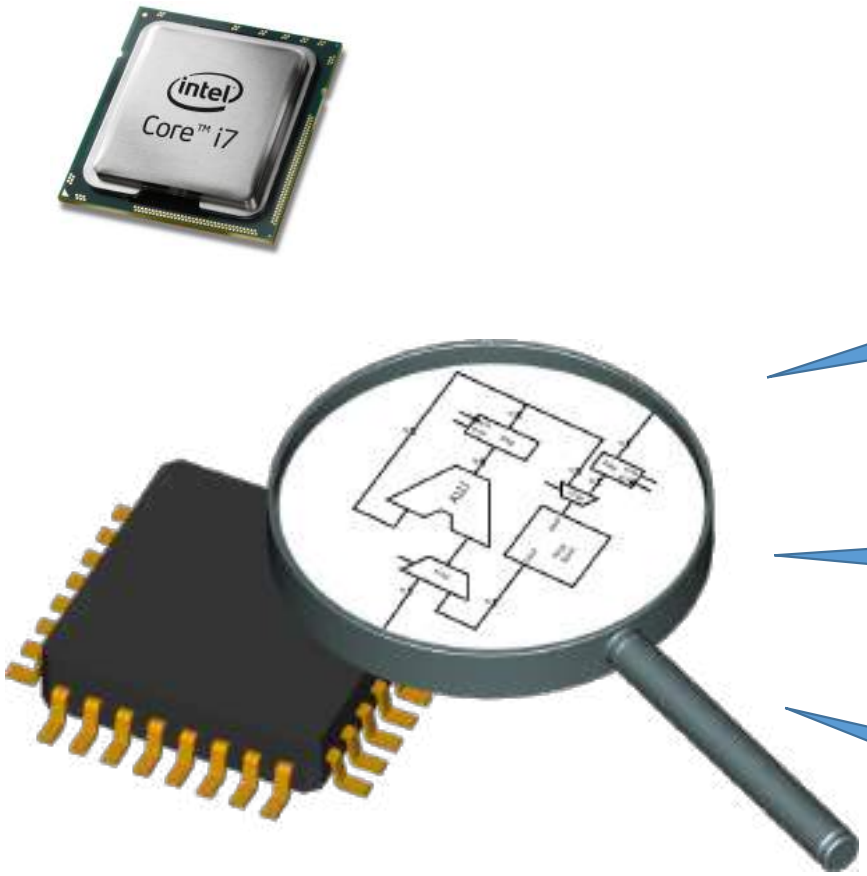
TensorFlow Lite




Cloud
Fog
= Fog Node

Nodos en la Niebla

Soluciones "*hardware*"



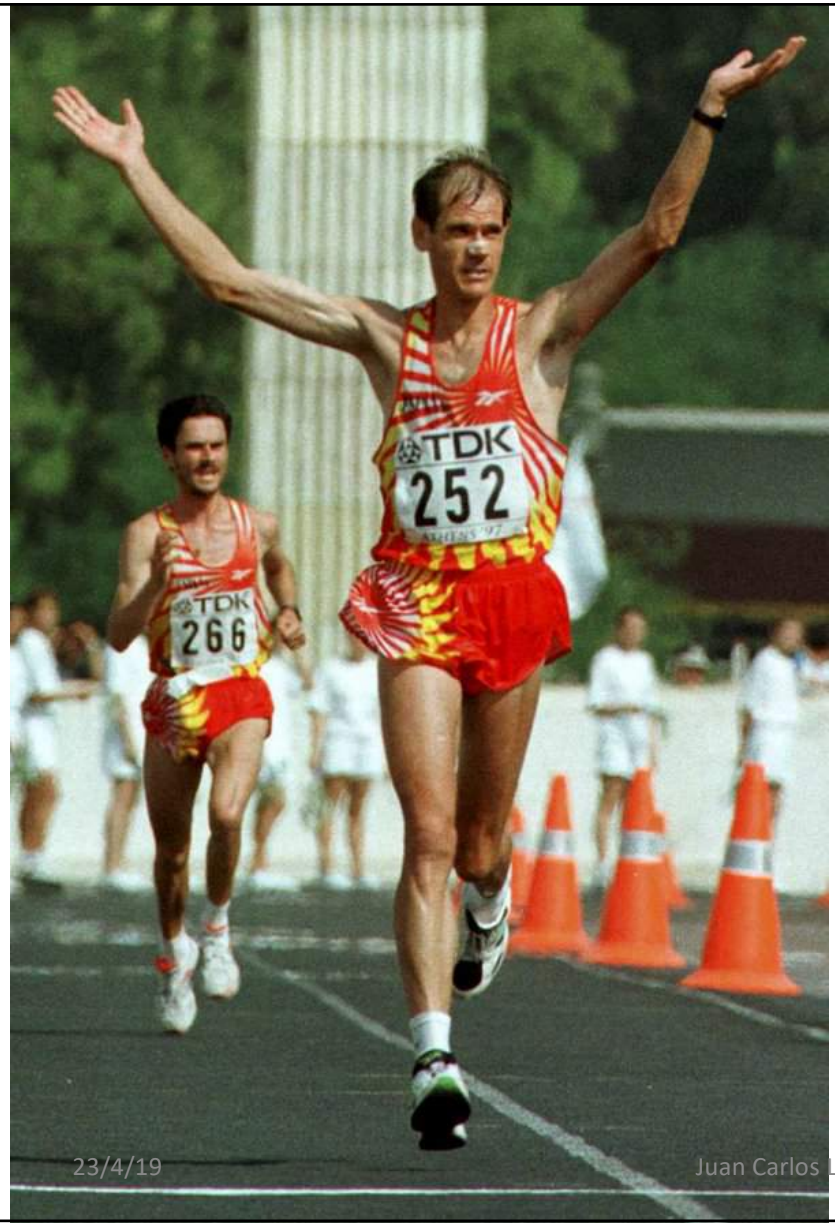
Programa para procesamiento de vídeo

Programa para encriptación de datos

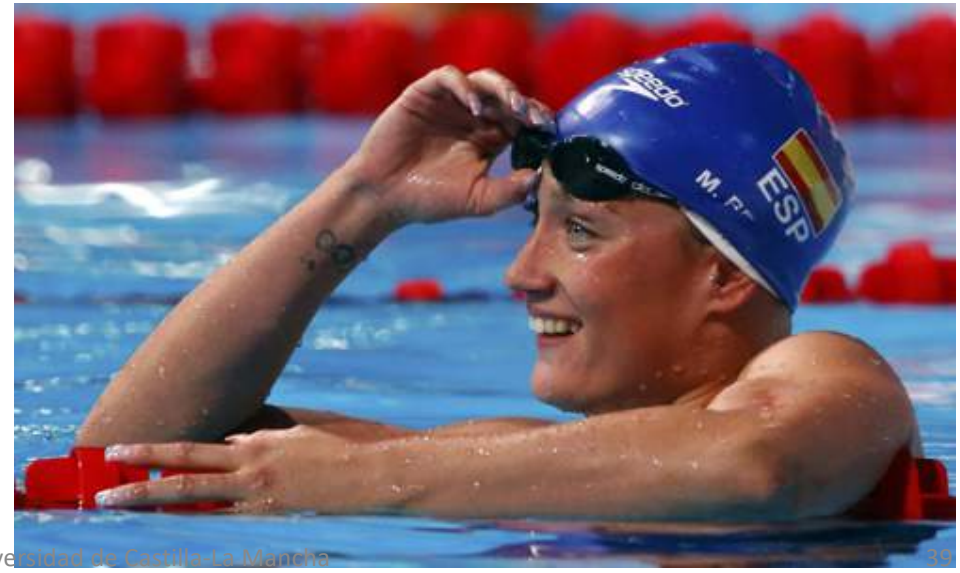
Programa para gestión de las comunicaciones


kiwami
MAVIC
TIME
artrouvillle
VIOLA

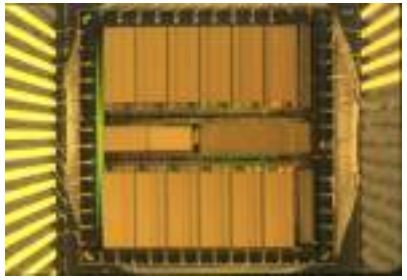

YOKOHAMA
ITU
WORLD TRIATHLON


TDK
252
TDK
266

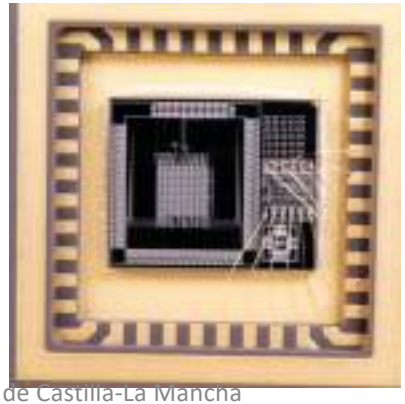

TOUR
LYONNAIS


ESP

Gráficos

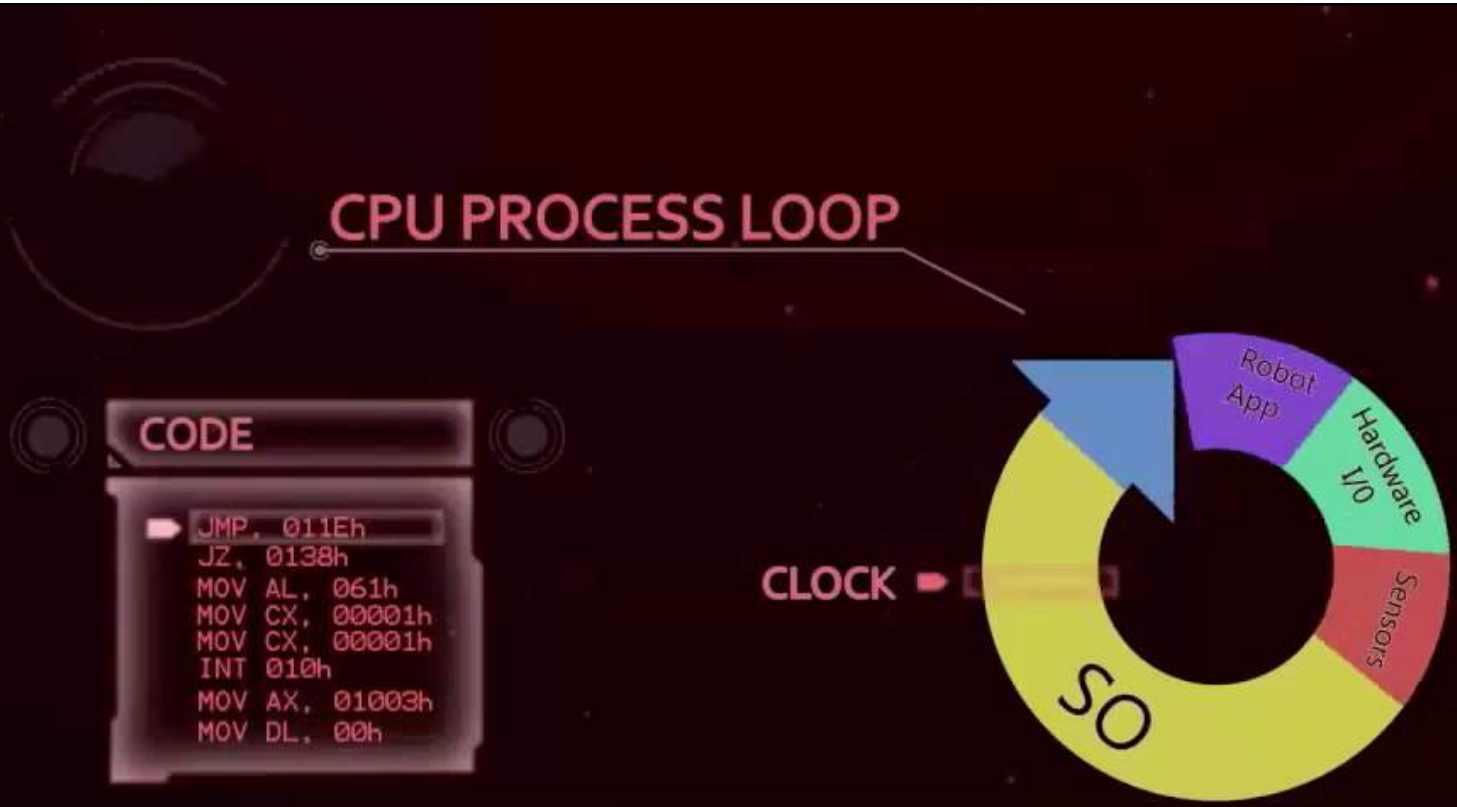


Criptografía



Comunicaciones

CPU PROCESS LOOP



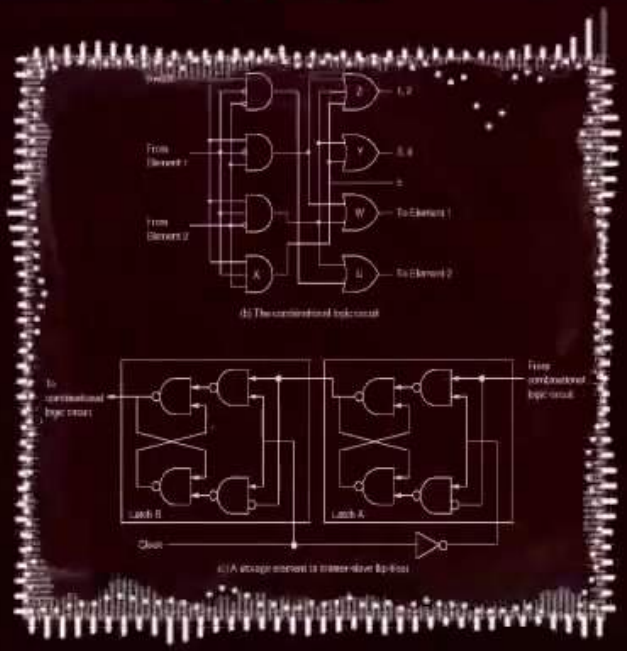
Más velocidad = Más consumo

CPU PROCESS LOOP

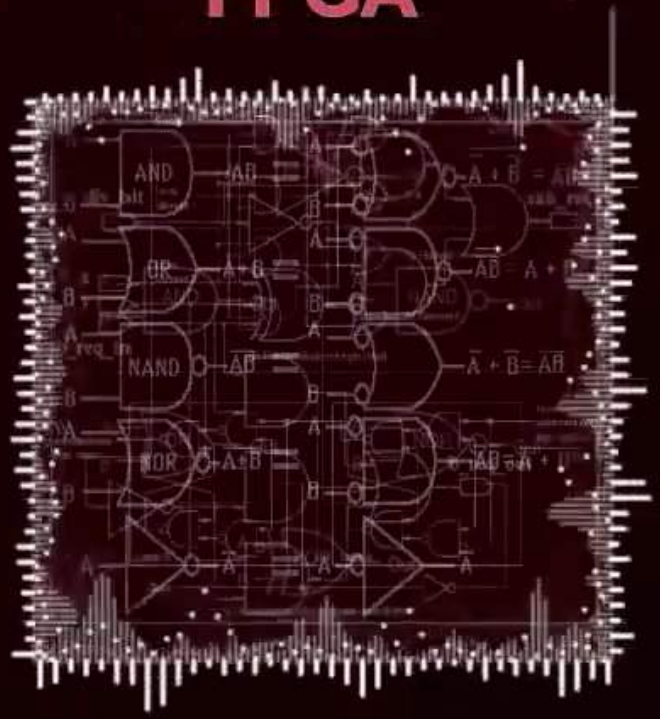

MAVIC
TIME
artrouville
VIOLA


02:03:06
Santander
123

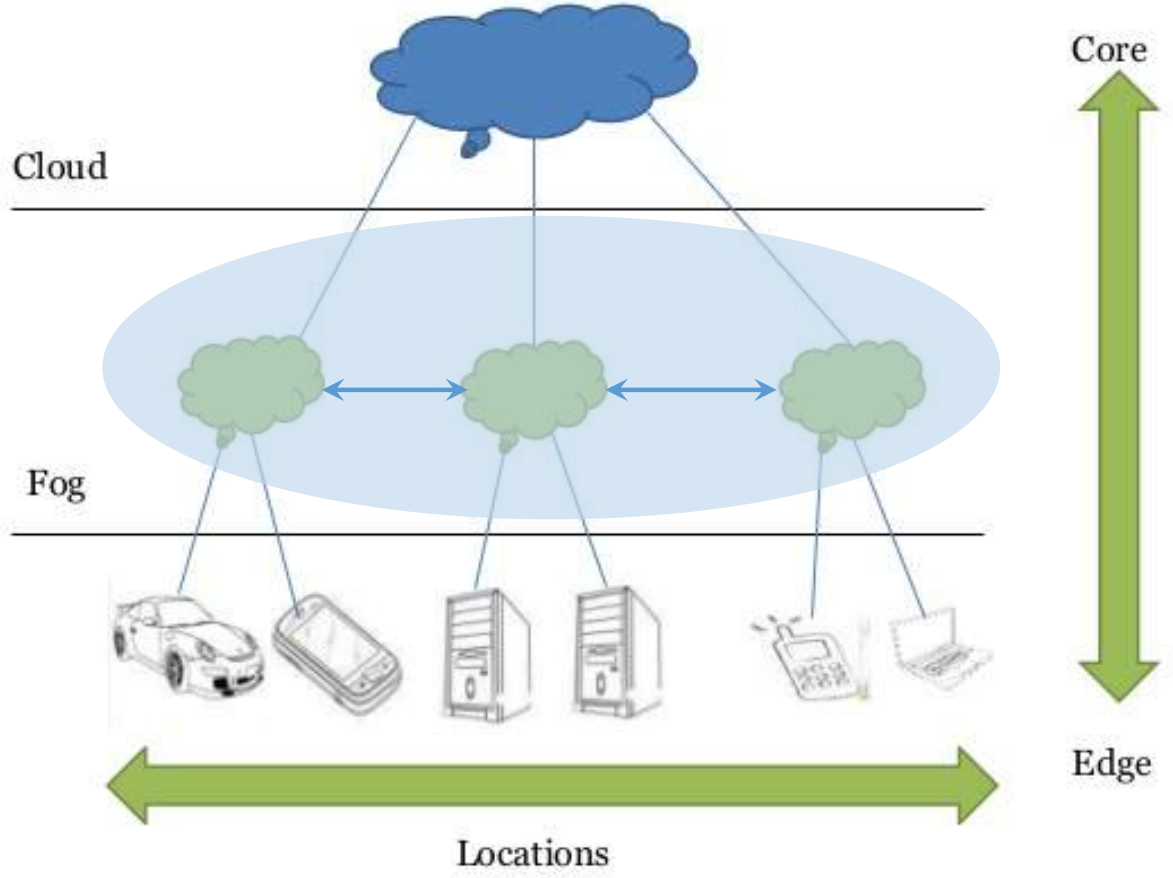
ASIC



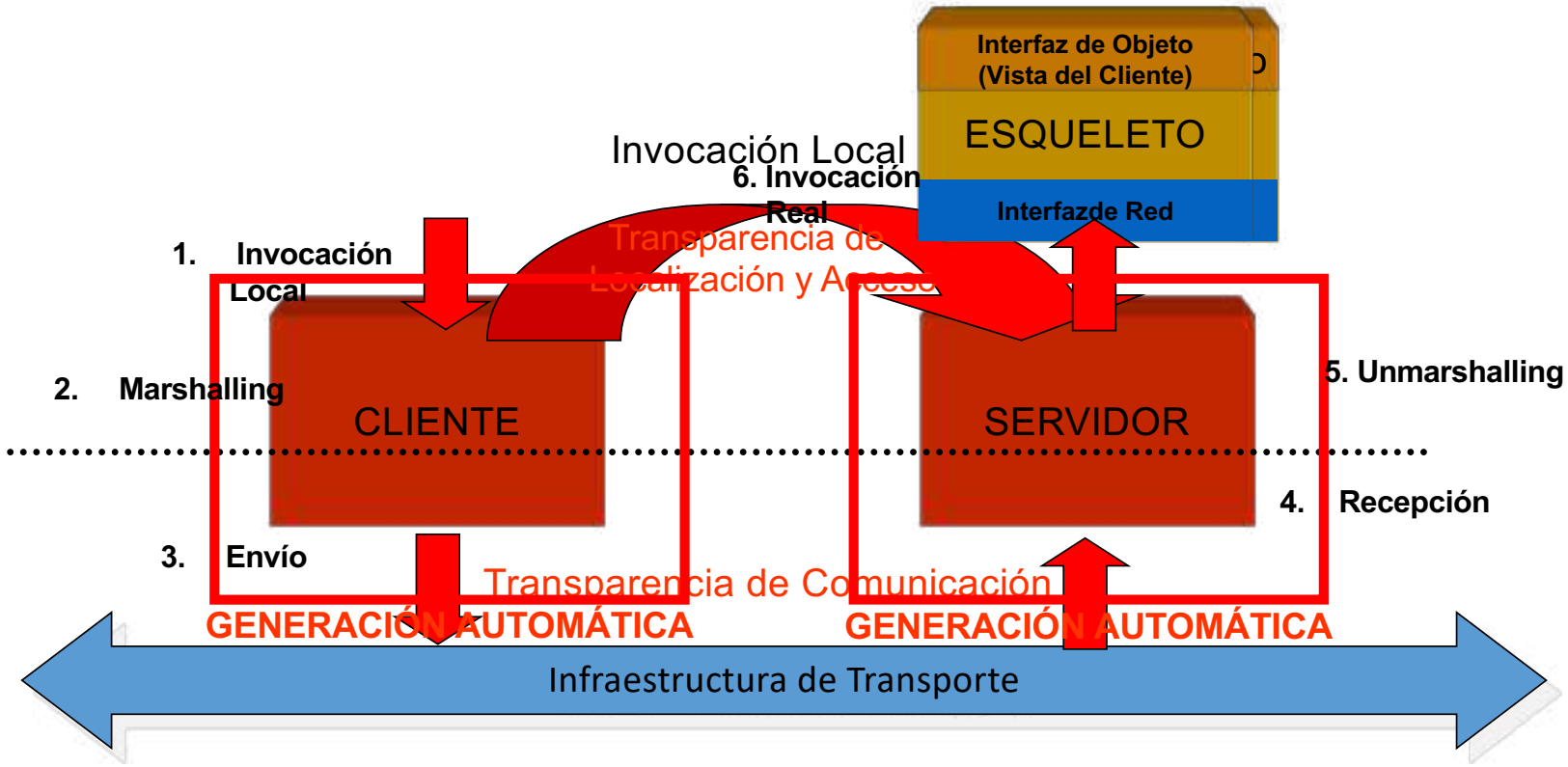
FPGA



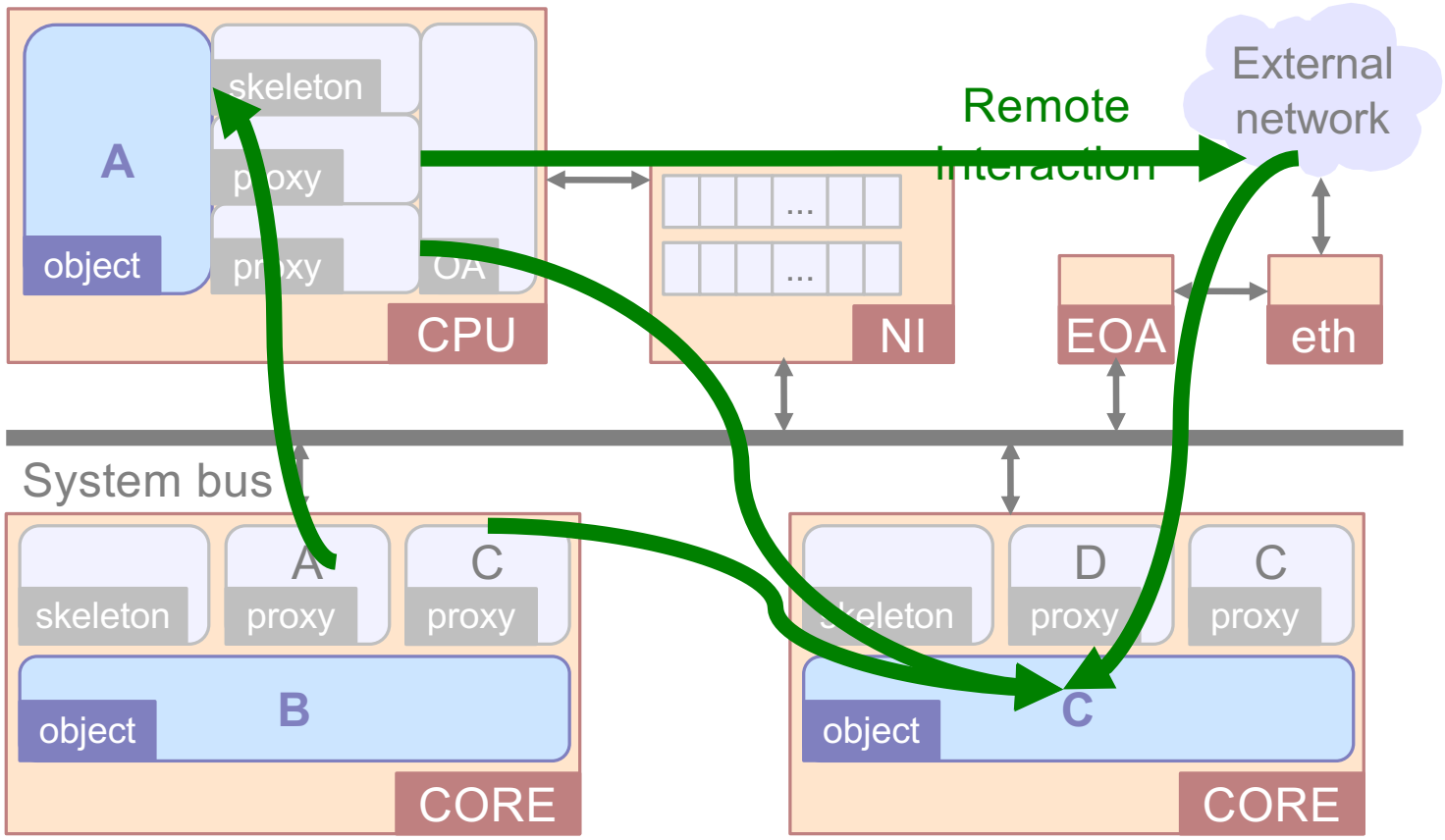
Conectividad *hardware*


Core
Cloud
Fog
Edge
Locations

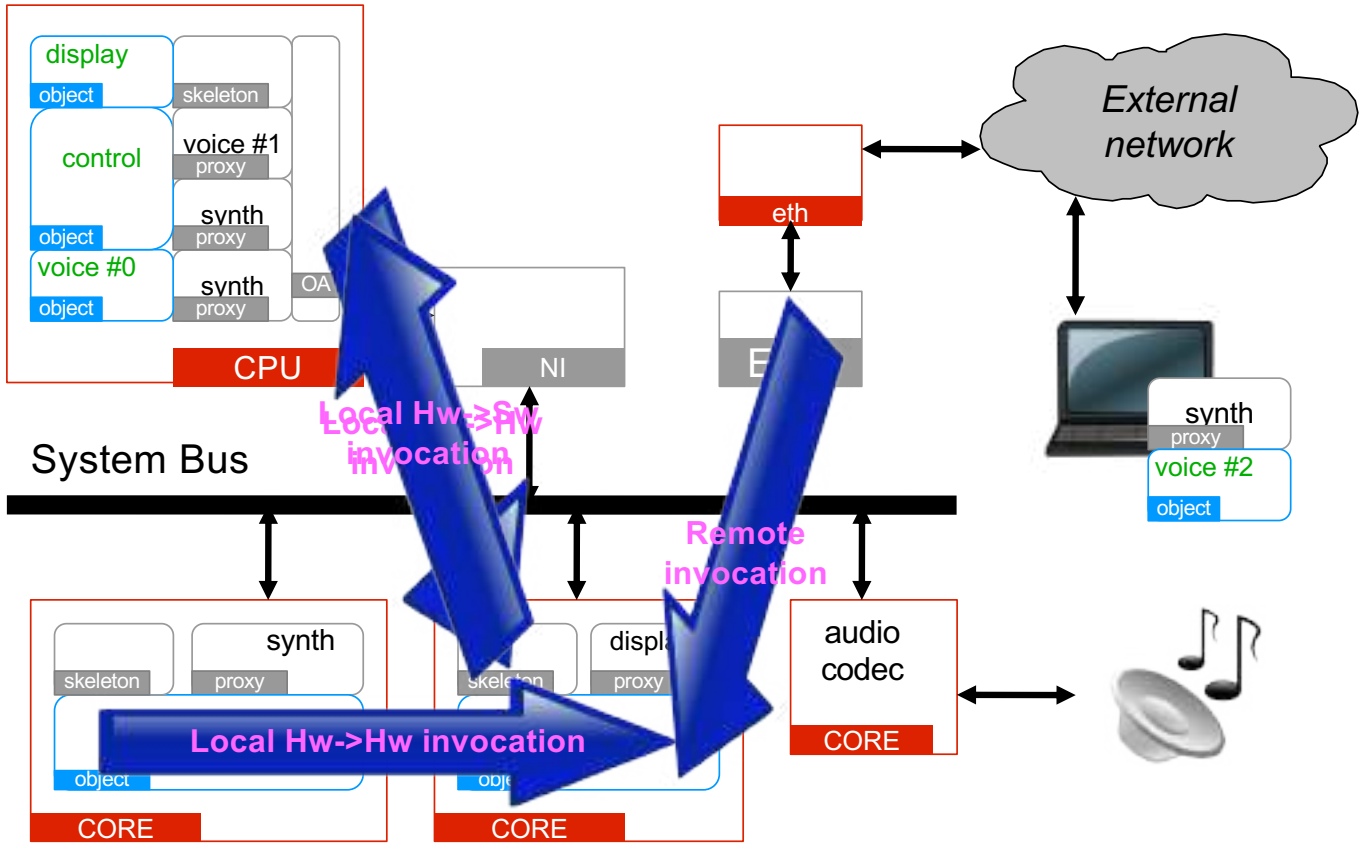
Invocación a Método Remoto (RMI)



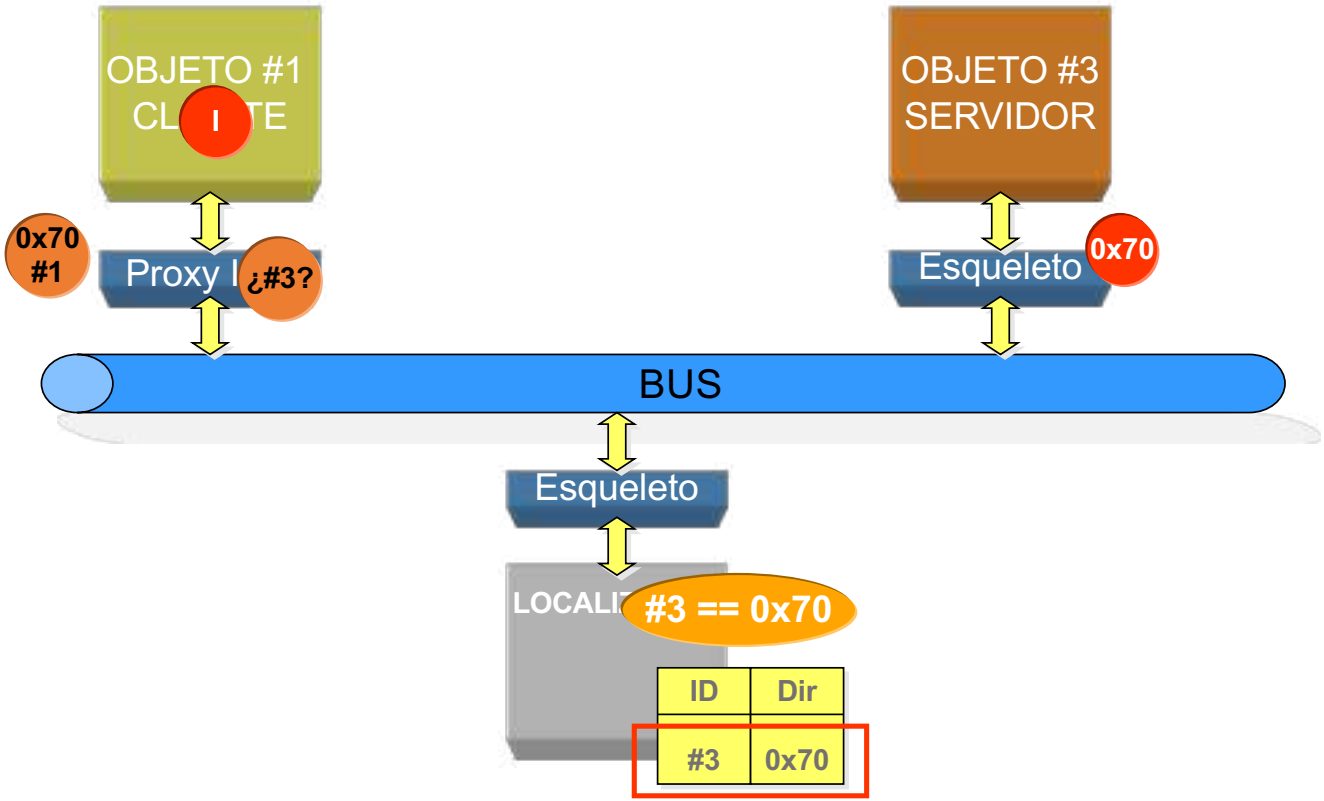
Arquitectura global de comunicaciones



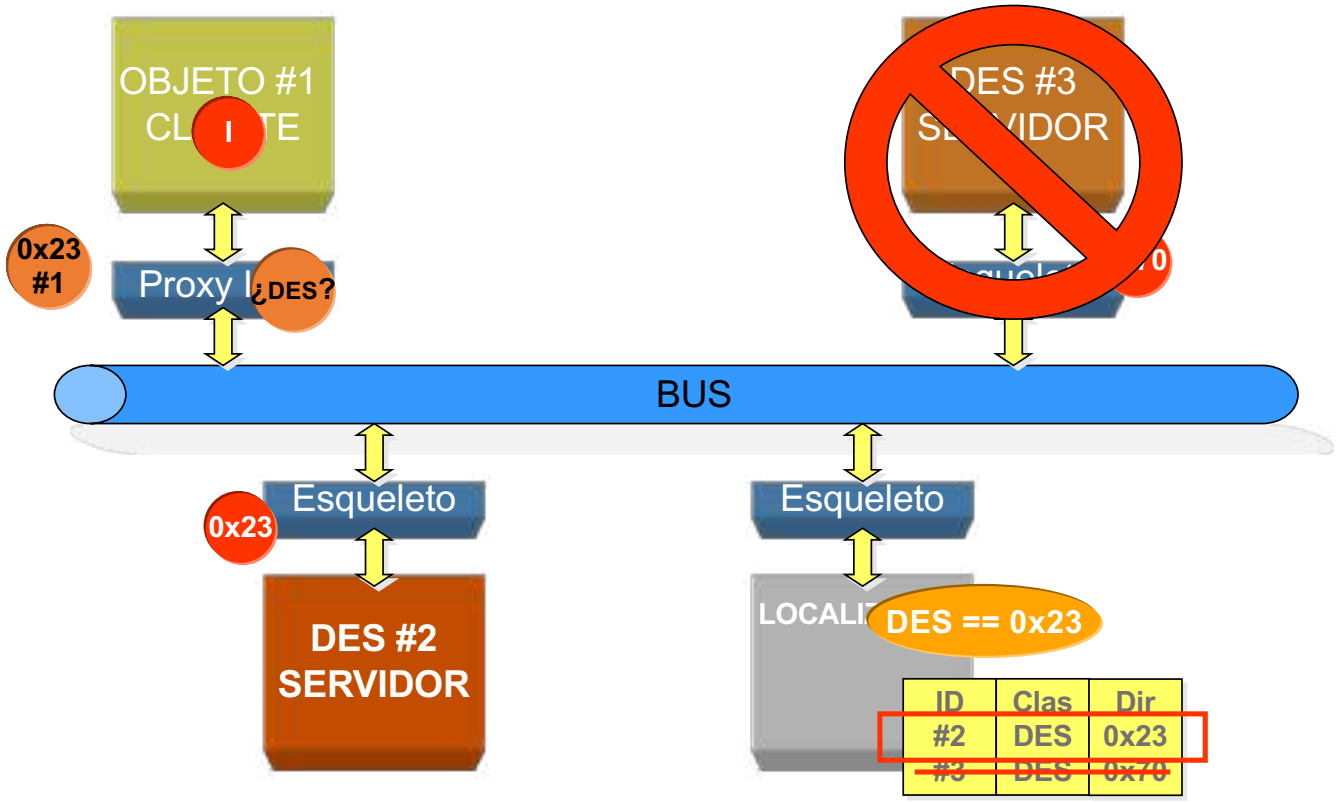
Middleware de sistema



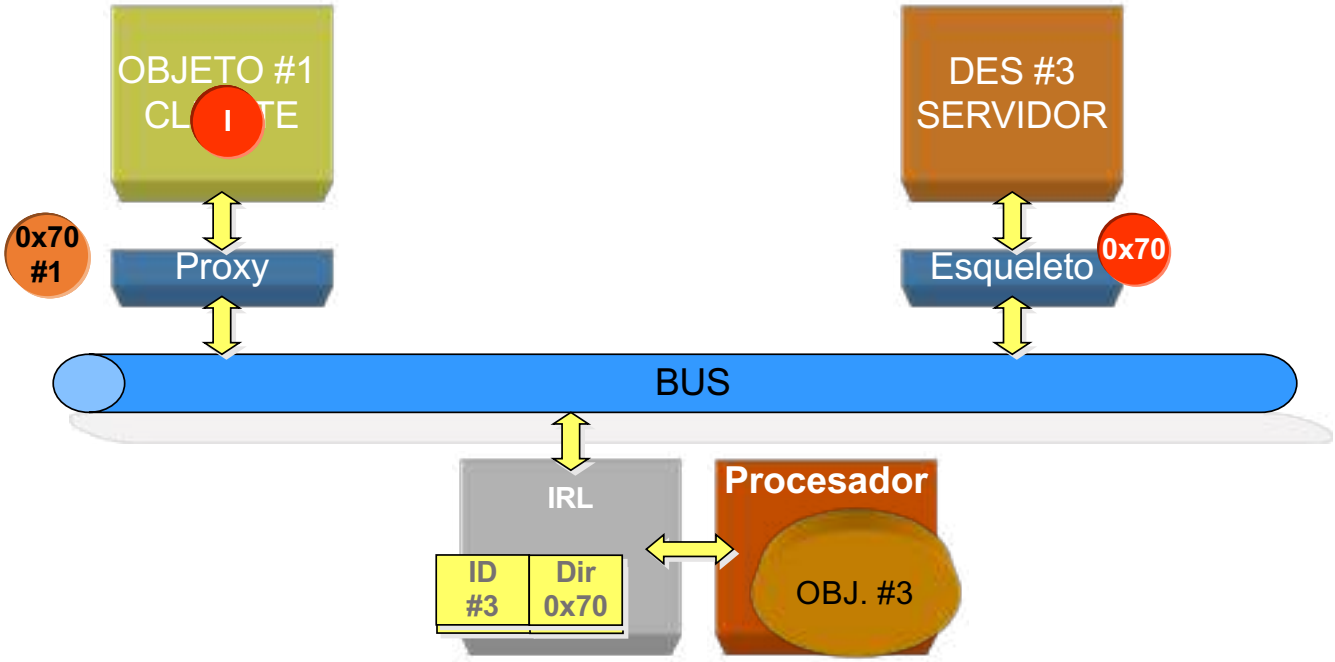
Servicio de localización



Aplicación: Tolerancia a fallos



Aplicación: Migración

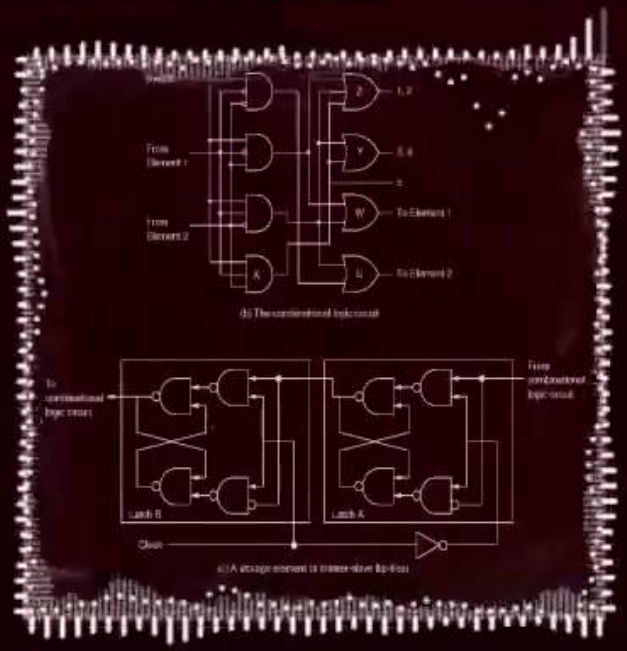


El papel de los elementos de la arquitectura

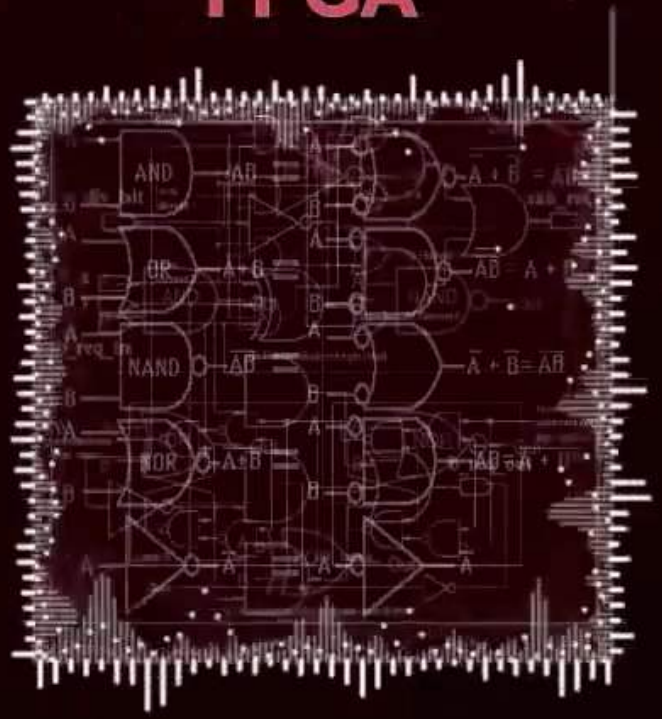
Componente	Concepto de OOCE	Aplicación en el diseño de SoCs
▪ Proxy ▪ Esqueleto	▪ Transparencia de acceso al canal ▪ Transparencia de acceso al componente	▪ Reutilización de componentes ▪ Independencia del canal de comunicación ▪ Intercambio de IPs
▪ Adaptador de objeto local	▪ Transparencia de localización	▪ Integración HW/SW ▪ Migración
▪ Interfaz de red local	▪ Transparencia de acceso al canal	▪ Independencia del canal de comunicación ▪ Migración HW/SW
▪ Adaptador de objeto remoto	▪ Transparencia de localización	▪ Comunicación fuera del chip
▪ Servicio de localización	▪ Transparencia de replicación, fallos, escalado, migración.	▪ Calidad de servicio ▪ Replicación ▪ Balanceo de carga
▪ Proxy indirecto ▪ Esqueleto indirecto	▪ Transparencia de migración, fallos, escalado, migración.	▪ Tolerancia a fallos ▪ Mejora de los tiempos de respuesta

Procesamiento/Prestaciones *hardware*

ASIC

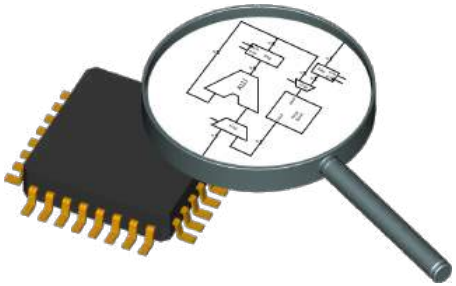


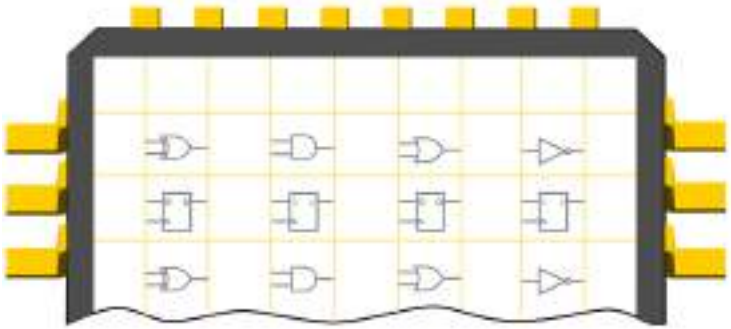
FPGA

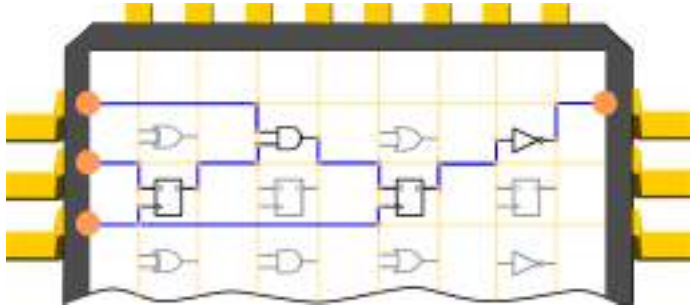


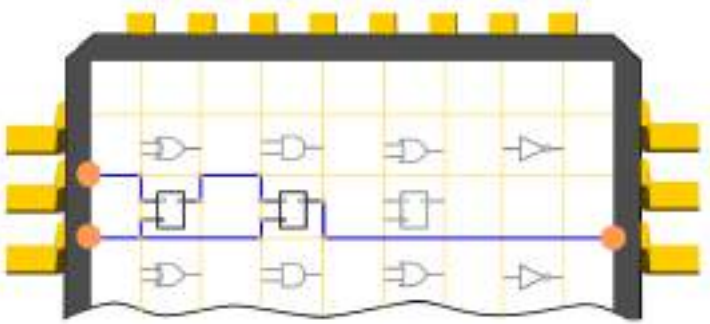
FPGA

Dispositivo Lógico Reconfigurable







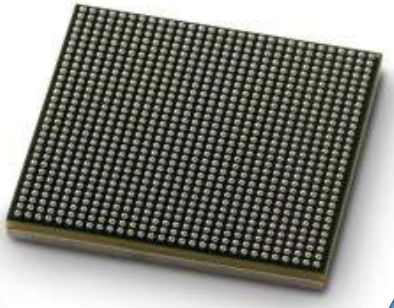


Gráficos

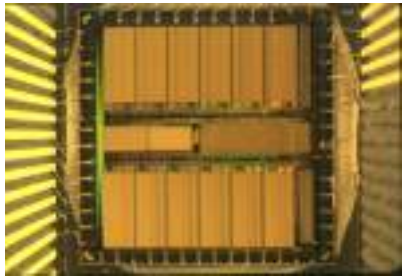
Criptografía


02:03:06
Santander
123

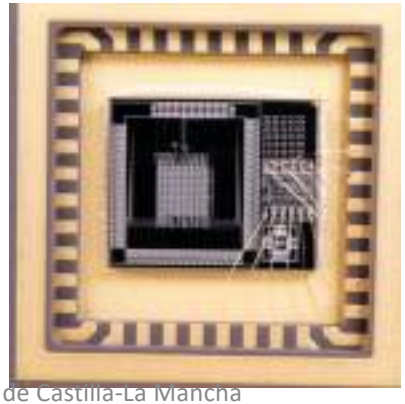
FPGA



Gráficos



Criptografía



Comunicaciones

Ya en Amazon y Microsoft

Futuros nodos en la niebla

FPGA para procesamiento

Pros

- Mejor ratio **coste / rendimiento**
- Flexibilidad (SW) y prestaciones (HW) - Paralelismo
- Menor **consumo**
  - FPGAs decenas de vatios (incluso mejor)
  - GPU cientos de vatios
- **Capacidad creciente**
- **Mejora** reconfiguración y herramientas

Cons

- Complejidad de diseño:
  - Dependencia de la tecnología
  - Compleja intercomunicación HW-SW
- Coste inicial elevado
- Sistemas empotrados (sin punto flotante eficiente)
- Adecuada para flujo de datos
- Dificultad memoria

Otras soluciones: GPUs

FPGA

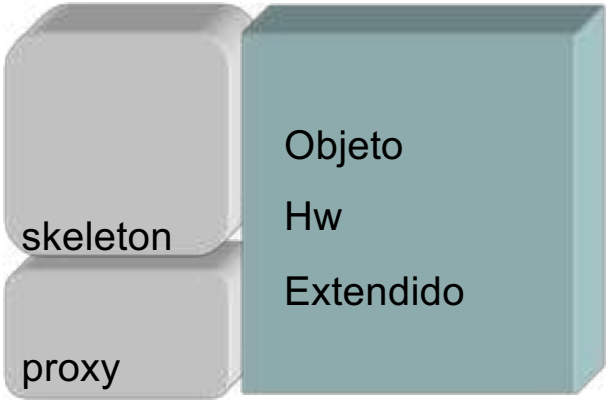
El problema de la reconfiguración

Dinámica

Parcial

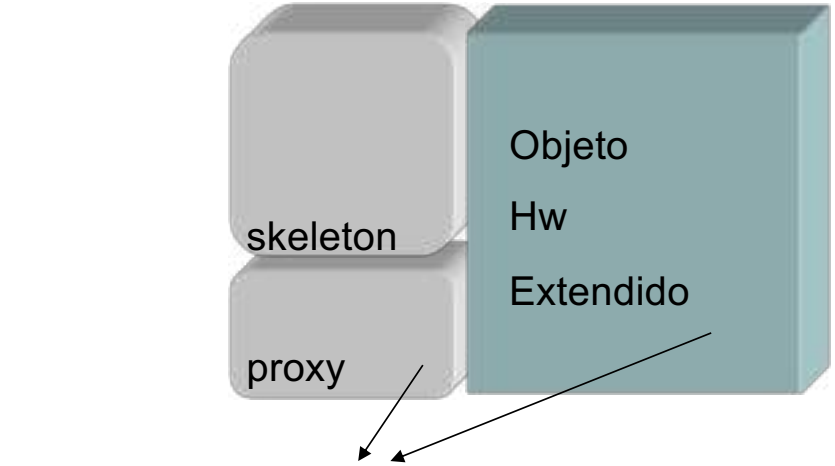
NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES



NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES

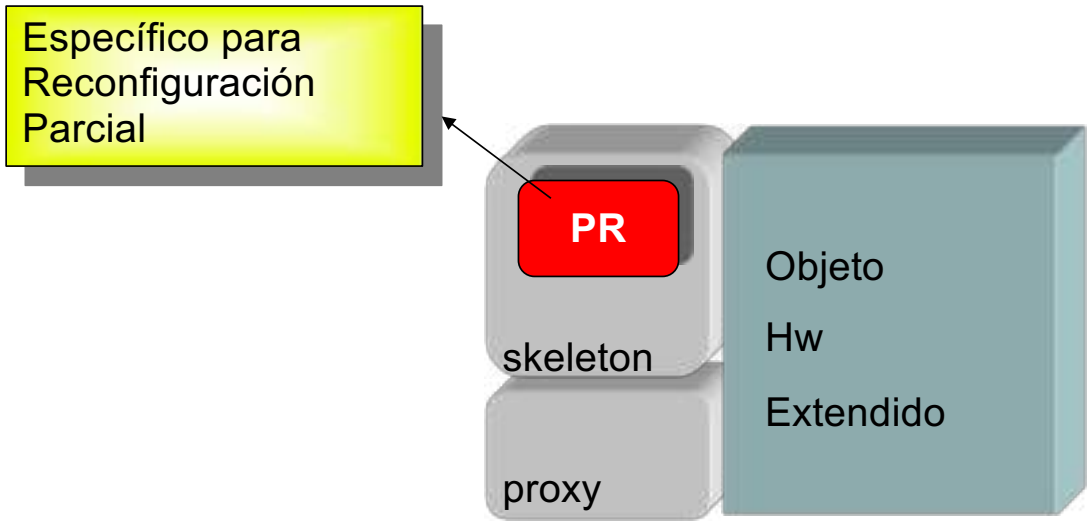


NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES

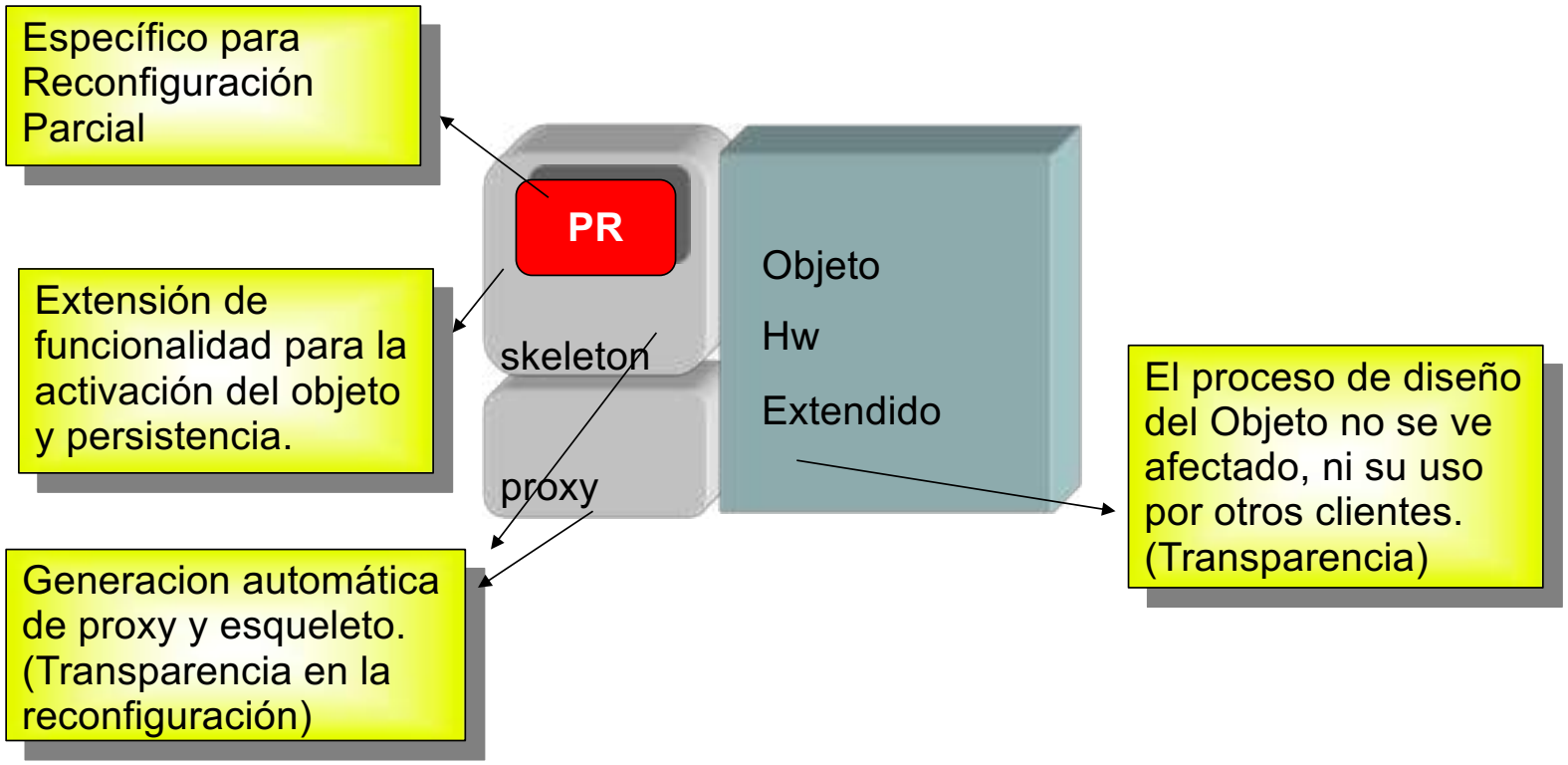
NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES



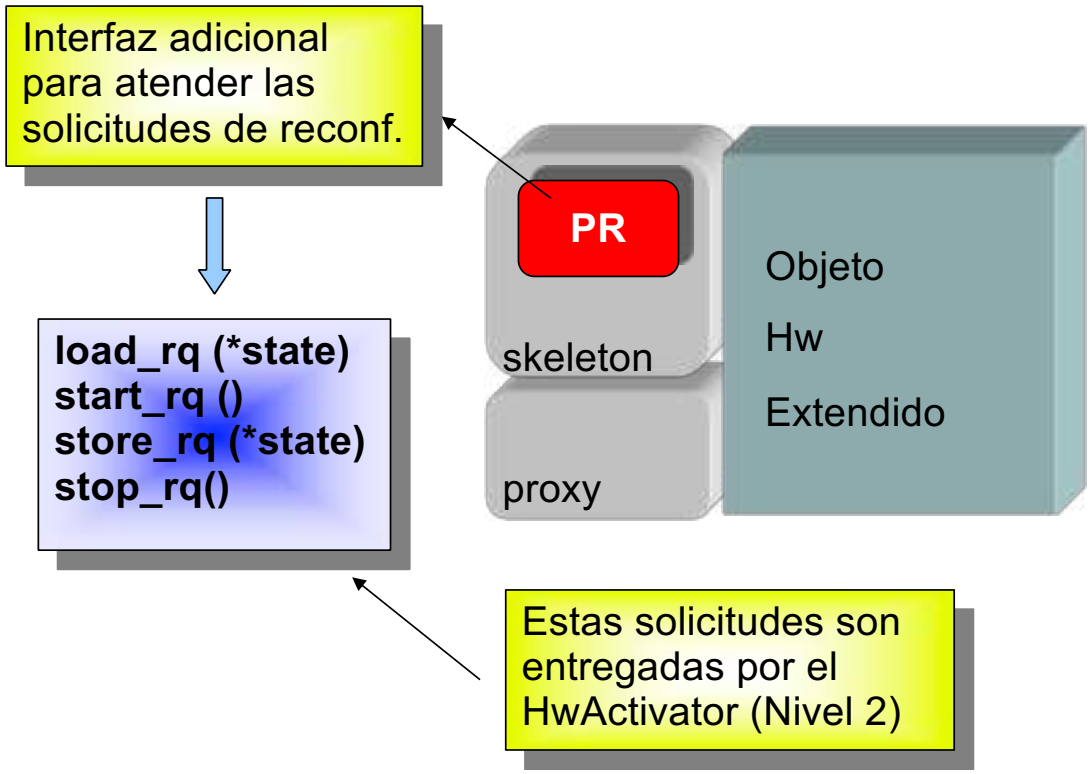
NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES



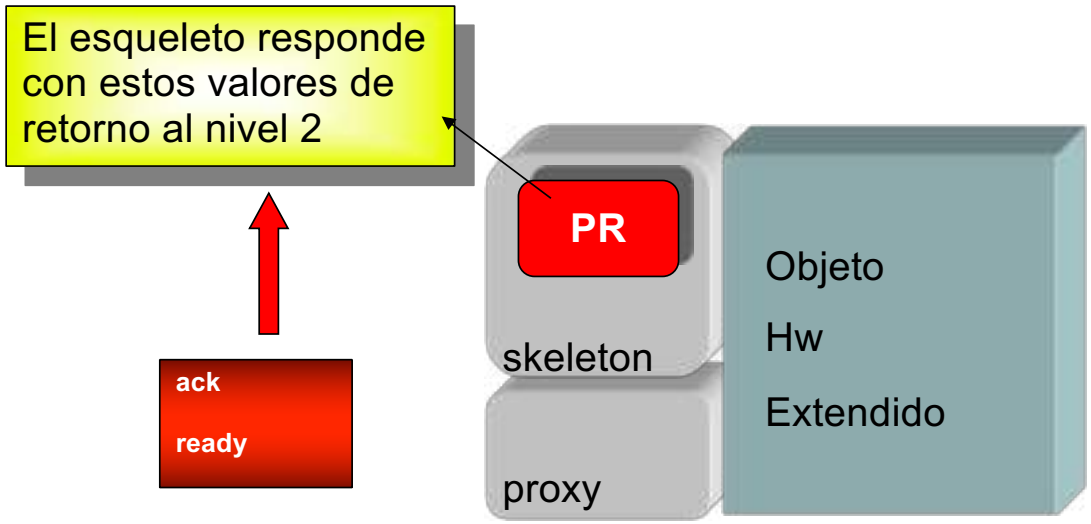
NIVEL 1

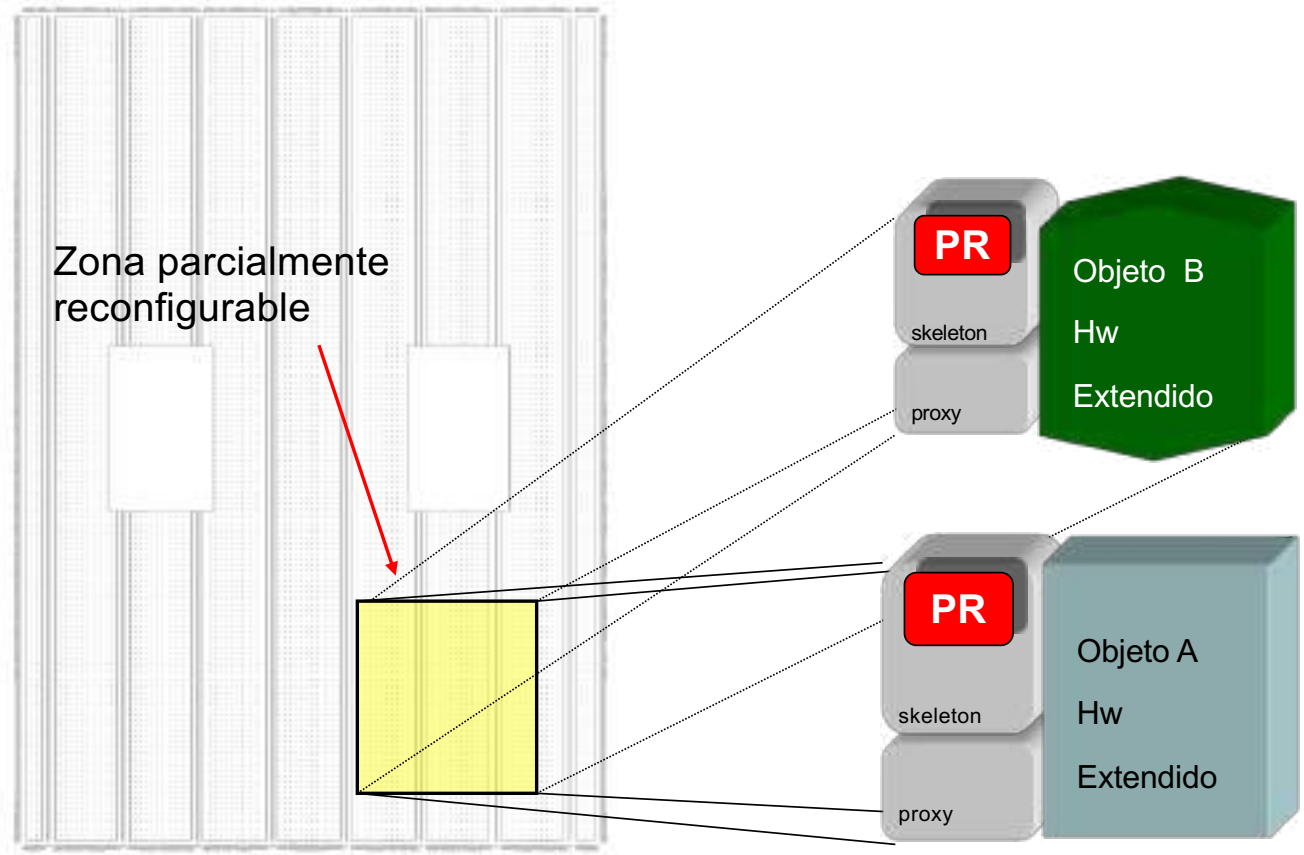
OBJETOS DINAMICAMENTE RECONFIGURABLES



NIVEL 1

OBJETOS DINAMICAMENTE RECONFIGURABLES




Zona parcialmente
reconfigurable
PR
skeleton
proxy
Objeto B
Hw
Extendido
PR
skeleton
proxy
Objeto A
Hw
Extendido

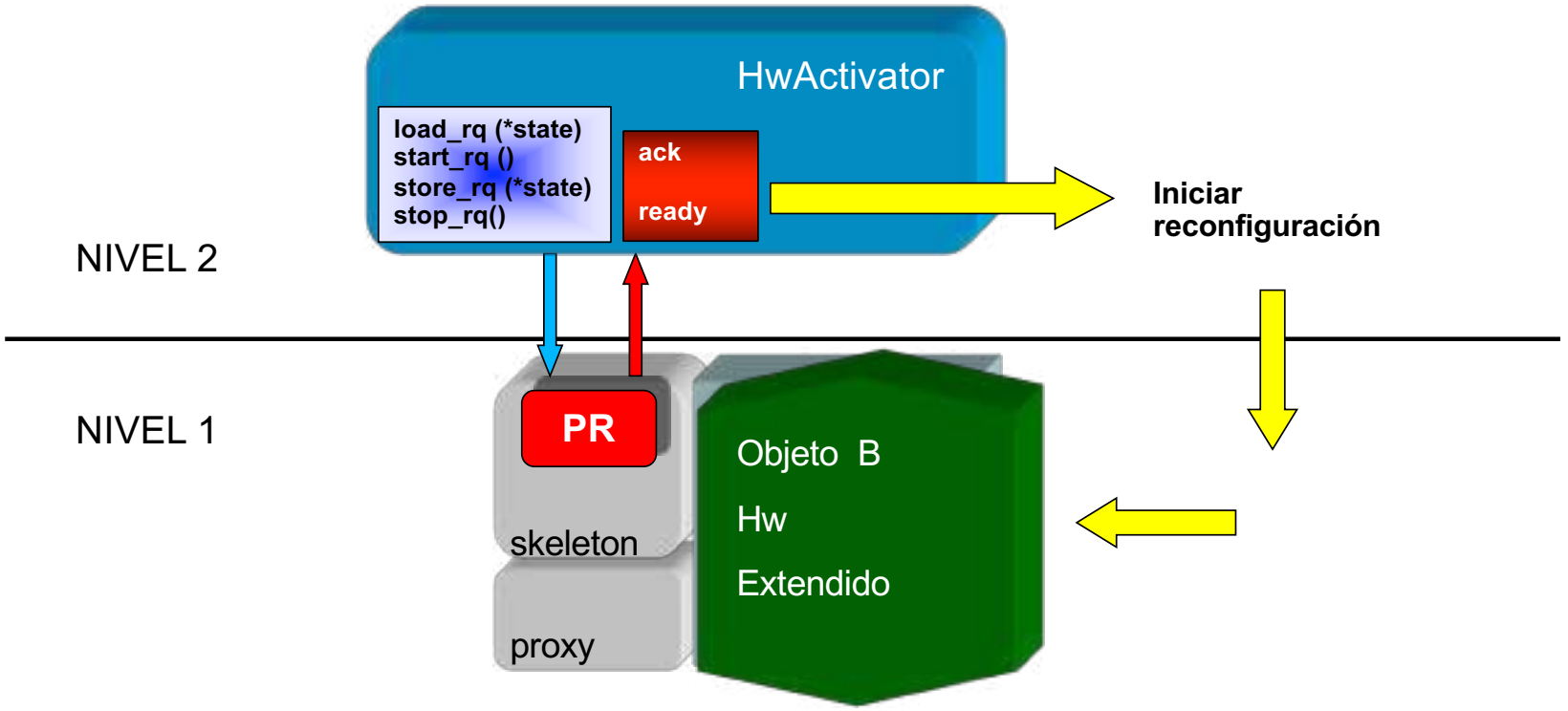
NIVEL 2
CAPA DE ACTIVACIÓN

Peticiones al NIVEL 1

Gestiona los servicios

- ❑ Parada de Objeto (stop_rq)
- ❑ Almacenamiento de estado (store_rq)
- ❑ Carga de estado (load_rq)
- ❑ Activación de Objeto (start_rq)

Secuencia de reconfiguración



NIVEL 3
CAPA DE PLANIFICACIÓN

Reconfiguration Manager:

Utiliza los servicios provistos por la CAPA DE ACTIVACIÓN

Tiene en cuenta el estado actual del sistema y las características de los objetos estáticos para proveer *scheduling* a nivel de sistema de los objetos dinámicos:

- Creación
- Reconfiguración
- Migración de objetos Hw/Sw

Locator:

Conoce la ubicación y el estado de todos los objetos dinámicamente reconfigurables.

NIVEL 3
CAPA DE PLANIFICACIÓN

Reconfiguration Manager:

- *Tabla de recursos*: captura las características de cada tipo de ODR

obj_type	: object type identifier
obj_size	: size (area) and shape of the object
state_size	: size in words of the object state
bitstream_loc	: size in KB of the partial bitstream

Tabla 1. Resource table fields

- *Allocator*: provee bloques de memoria contiguos, para el almacenamiento del estado o de los bitstreams de reconfiguración.

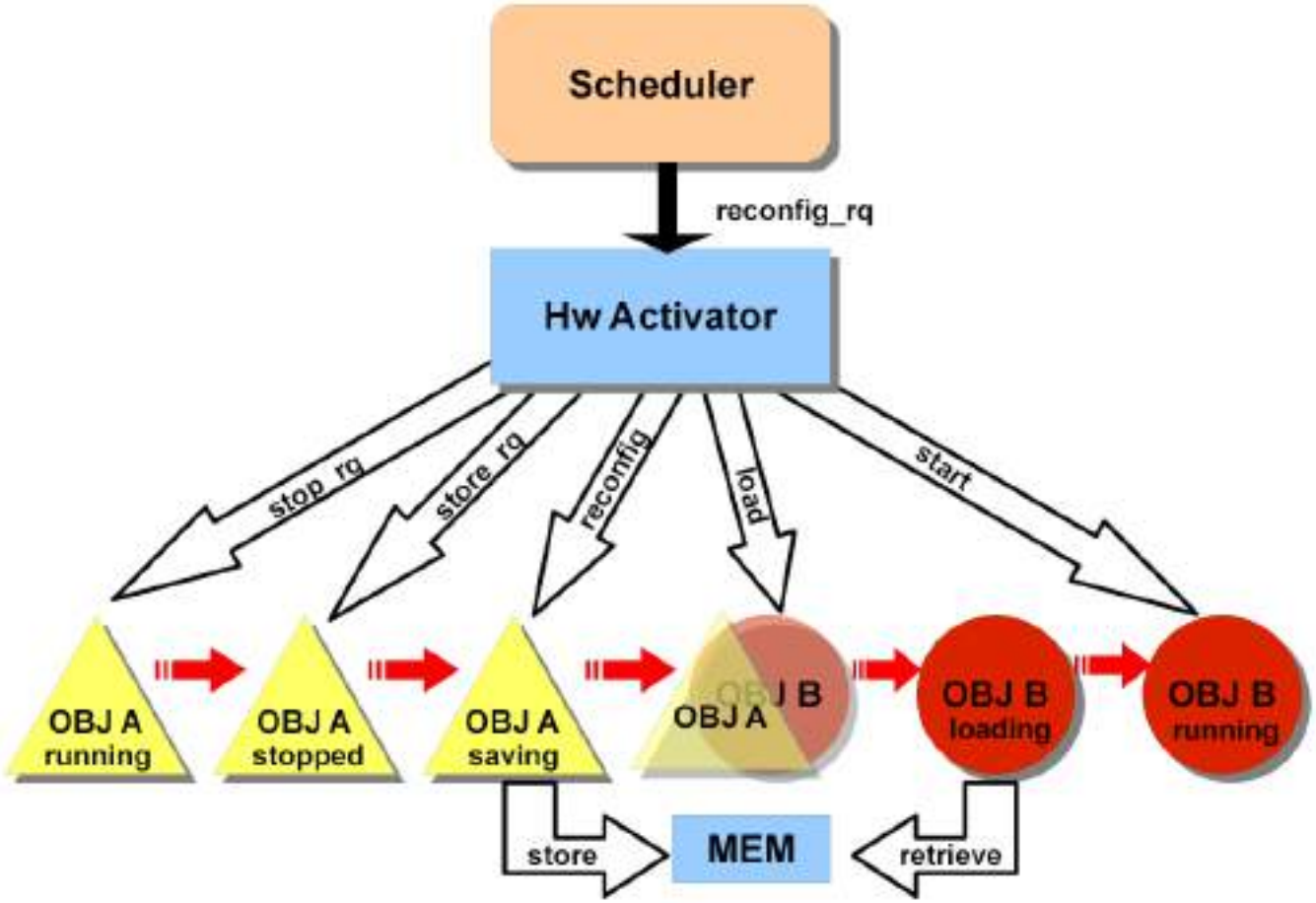
- *Scheduler*: toma decisiones respecto a la ubicación de nuevos objetos, la reconfiguración de objetos existentes y la migración sw/hw o viceversa.

NIVEL 3
CAPA DE PLANIFICACIÓN

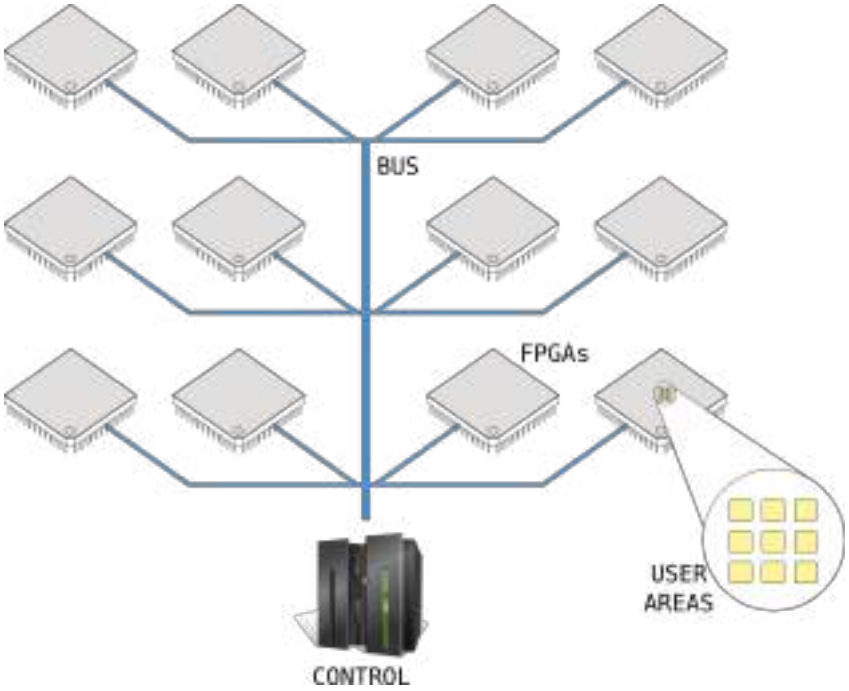
Locator: posee la información de la situación actual de todos los objetos dinámicos del sistema. Esta información es almacenada en una tabla de ubicación

```
obj_id          : object identifier
location        : object placement
                  (PR area/Sw processor)
status          : running, stopped or frozen
implementation  : hardware or software
storage         : memory space assigned for persistent
                  storage
```

Provee la ubicación real de los objetos dinámicos a cualquier otro objeto que necesite invocarlo. De esta manera los objetos clientes pueden acceder de manera transparente a objetos dinámicos. Permite creación implícita y explícita de objetos.


Scheduler
reconfig_rq
Hw Activator
stop_rq
store_rq
reconfig
load
start
OBJ A
running
OBJ A
stopped
OBJ A
saving
OBJ A
OBJ B
OBJ B
loading
OBJ B
running
store
MEM
retrieve

¿Cómo pueden integrarse las FPGAs en un sistema de procesamiento HPC?



- **Abstracción** del hardware
- **Modelo** de programación orientado a sistemas de objetos distribuidos

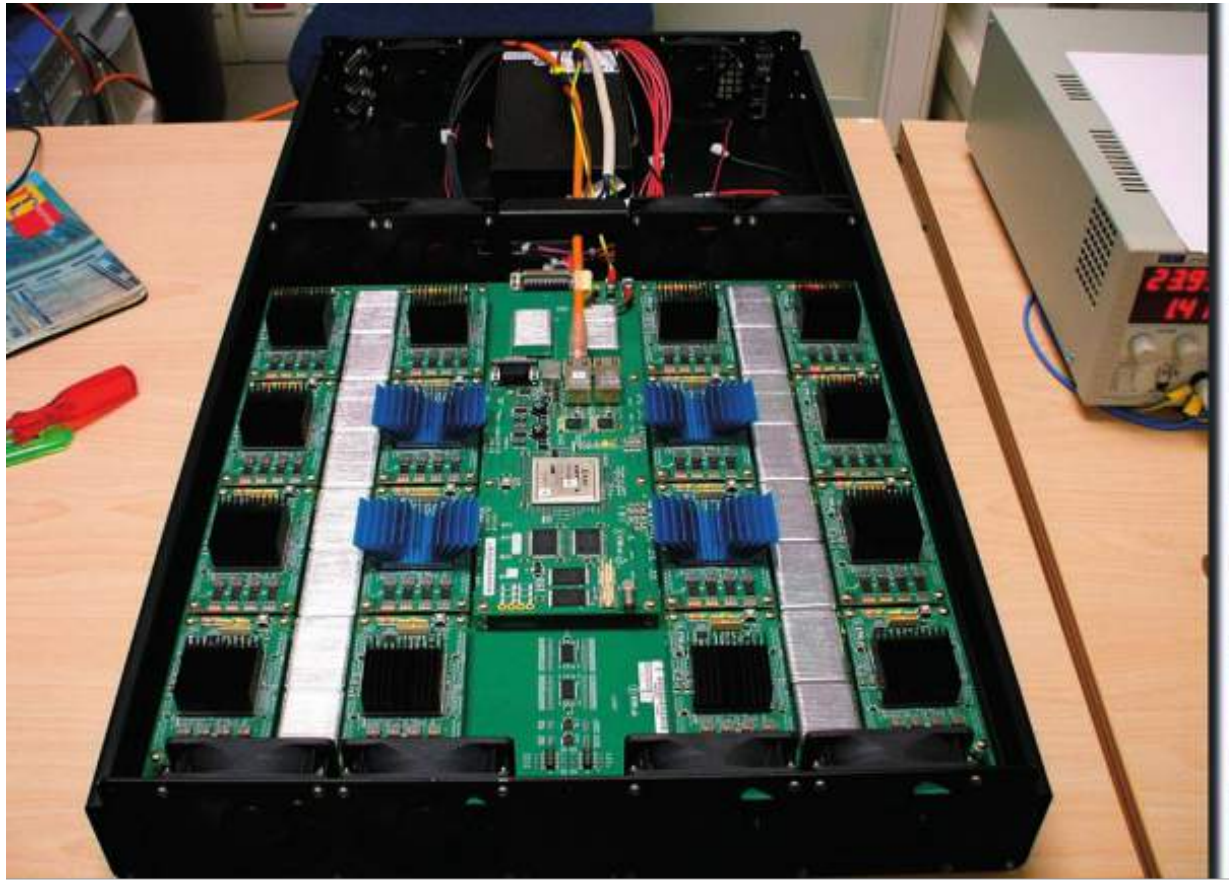
¿Cómo pueden integrarse las FPGAs en un sistema de procesamiento HPC?

Recurso computacional

Facilitar modelado del problema

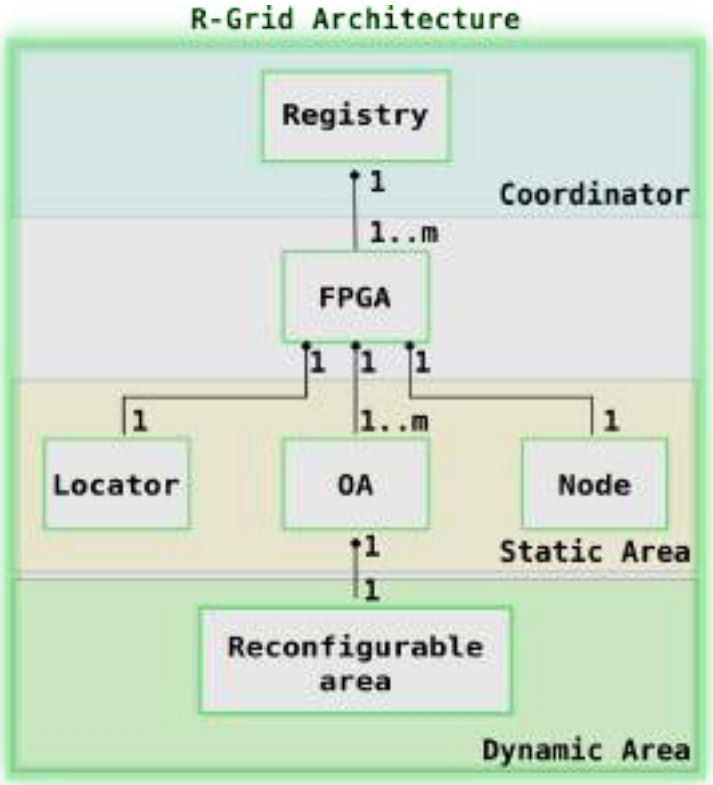
Despliegue automático

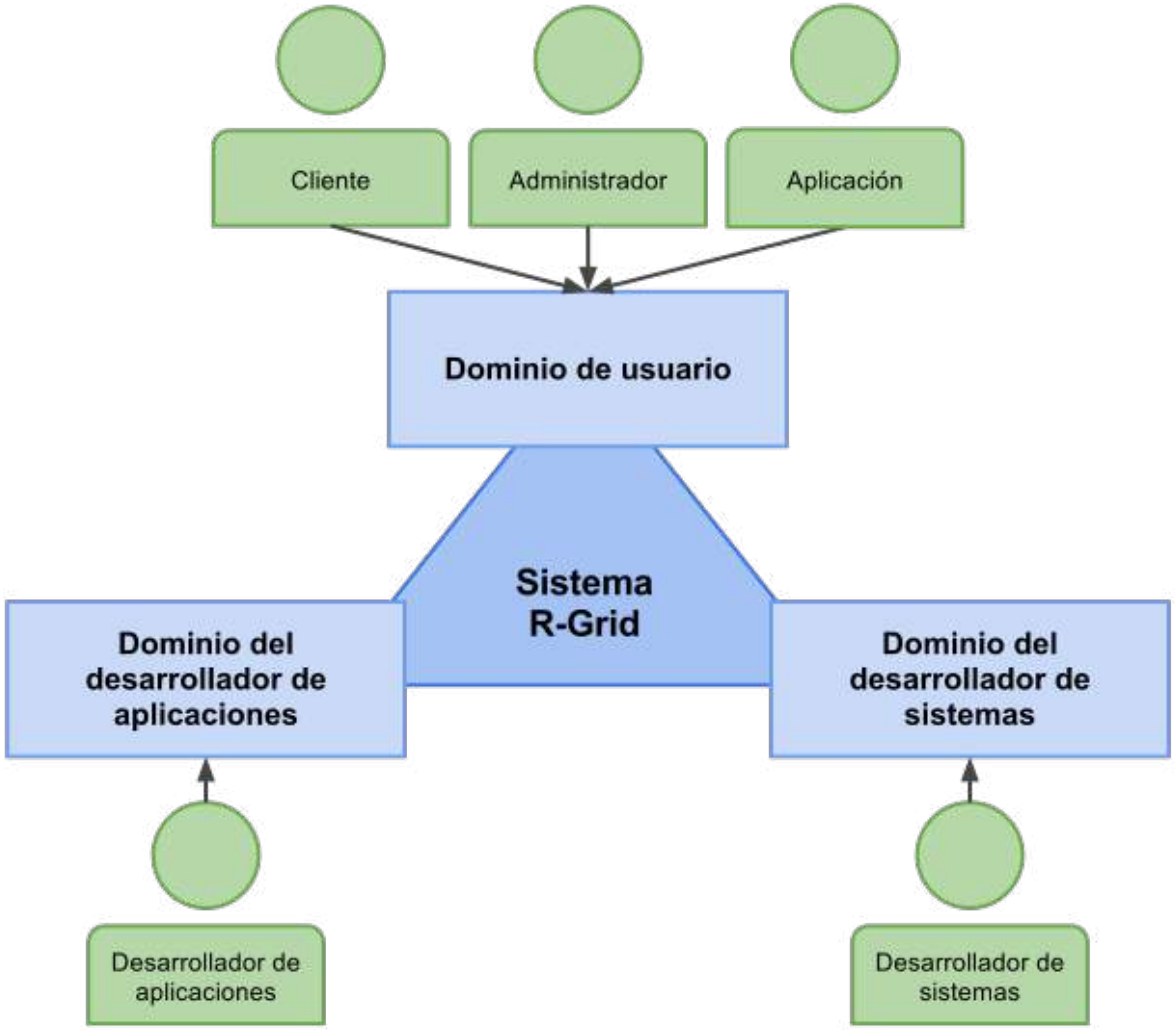
- Transparencia de localización
- Transparencia de comunicación
- Replicación y migración

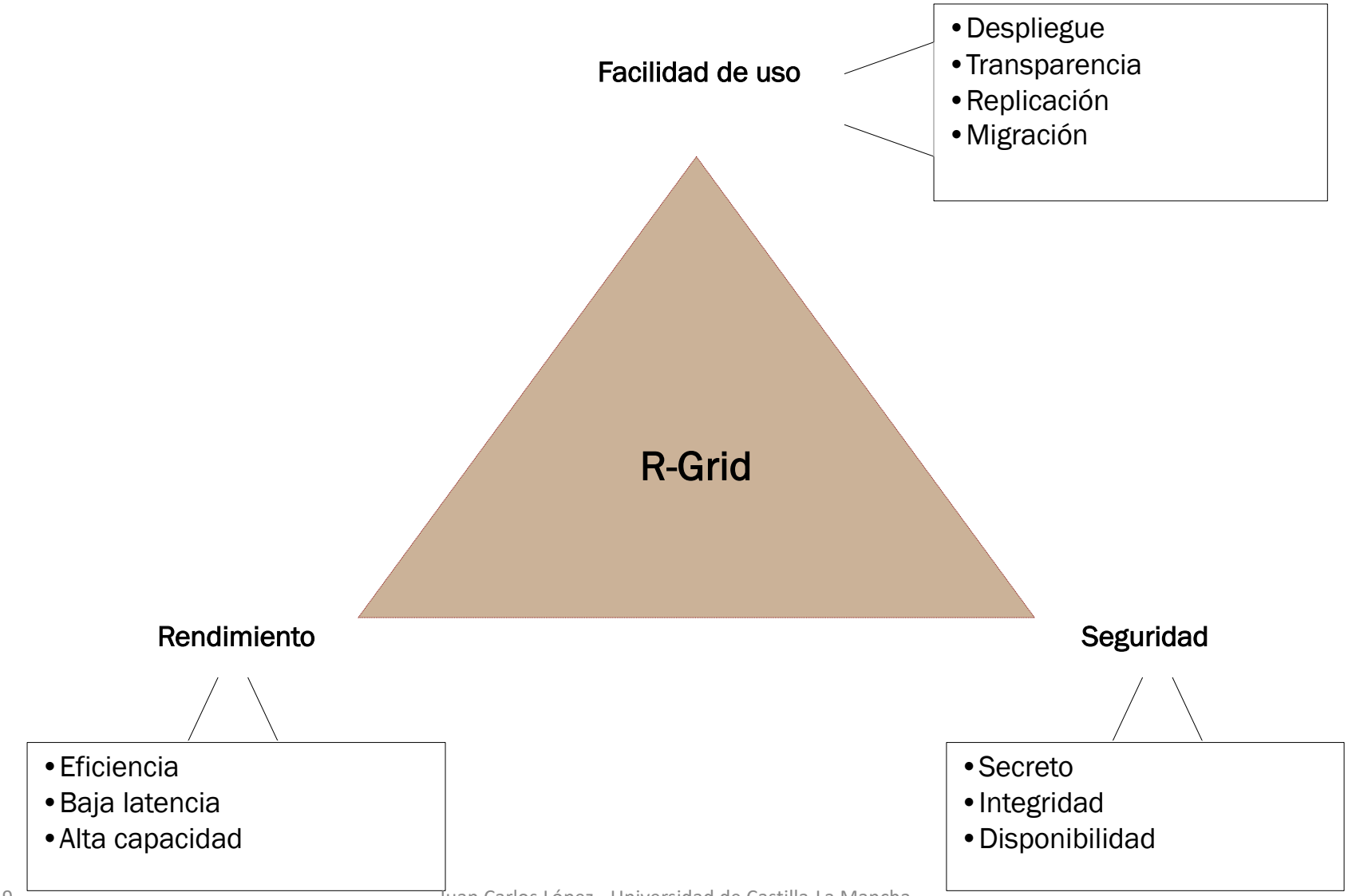


¿Cómo se gestiona?

- **R-Grid** posee tres niveles:
  - Coordinación R-Grid **(Coordinator)**
  - Capa de abstracción recursos Hardware **(Static Area)**
  - Recursos de usuario **(Dynamic Area)**




Cliente
Administrador
Aplicación
Dominio de usuario
Sistema
R-Grid
Dominio del
desarrollador de
aplicaciones
Dominio del
desarrollador de
sistemas
Desarrollador de
aplicaciones
Desarrollador de
sistemas


Facilidad de uso
• Despliegue
• Transparencia
• Replicación
• Migración
R-Grid
Rendimiento
• Eficiencia
• Baja latencia
• Alta capacidad
Seguridad
• Secreto
• Integridad
• Disponibilidad

FPGAs – *Fog Computing*
Aplicaciones

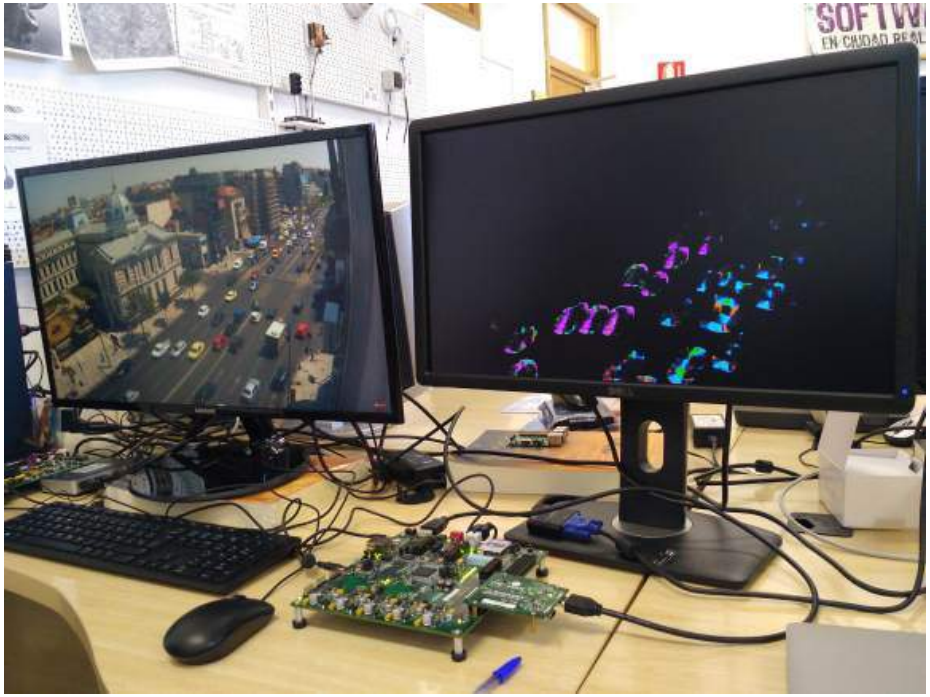
Procesamiento - Algoritmos

Imagen y vídeo

Sistemas empotrados de alto rendimiento (SEA)

High Performance Embedded Computing Systems

- Diseño e implementación de hardware a medida.
- Tecnología de hardware reconfigurable (FPGA):
  - Coste y tiempo de desarrollo reducidos (HLS).
- Soluciones de alto rendimiento, adaptables y bajo consumo.



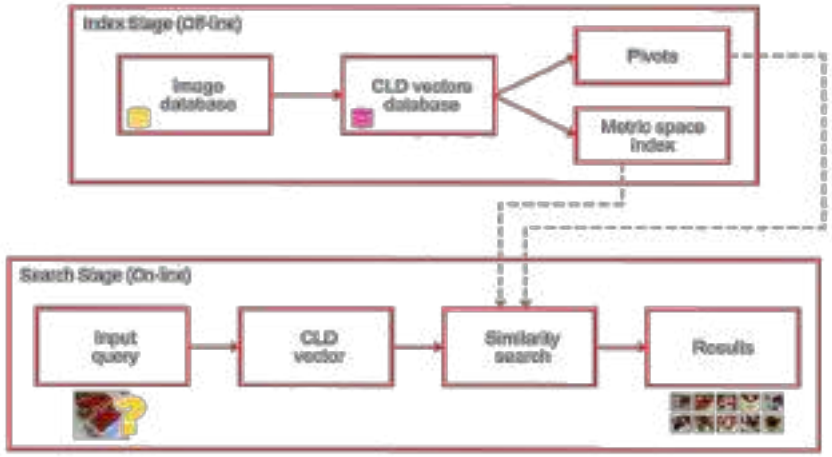
Análisis de flujo óptico (Lucas Kanade) en tiempo real (60fps) de vídeo en resolución FHD.

Detección de personas/animales

- Prototipo algoritmo HOG + SVM
- Mínimo uso de memoria
- Procesamiento en tiempo real (2 ciclos/pixel)

Otros

- Búsqueda de imágenes (MPEG-7) similares.
- Estimación de movimiento: implementación de un algoritmo de búsqueda de macrobloques
  - Parametrizable (resolución, tamaño de MB y área de búsqueda, puntos, etc.)
  - Mínima latencia (2 ciclos por pixel)
  - Reducido uso de memoria

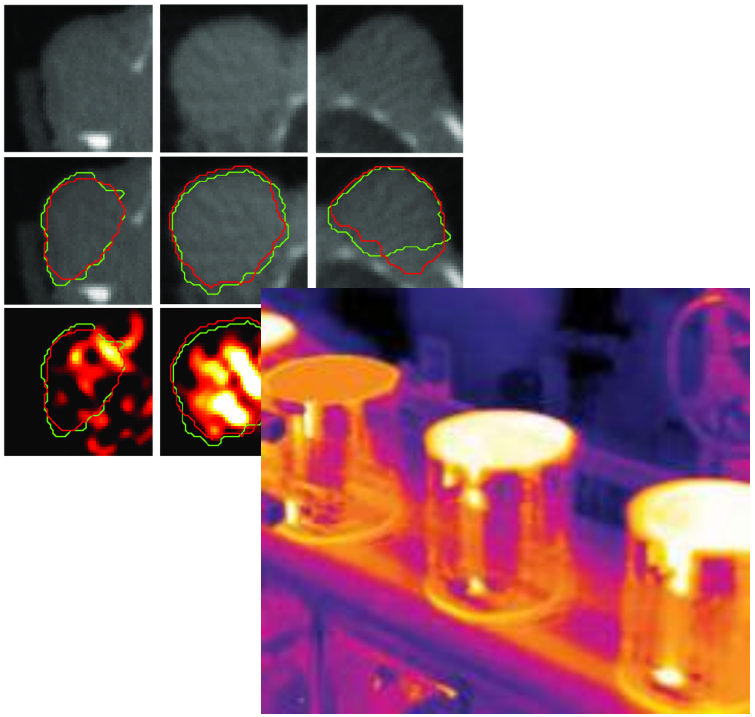


Aplicaciones de los SEA

Casos de uso desarrollados por ARCO

Alto rendimiento

- Aplicaciones de tiempo real estricto.
- Simulación.
- Aplicaciones científicas.
- Análisis de flujos de datos masivos.



Monitorización de procesos industriales, aseguramiento de la **calidad** (TRAZAVIN)

Asistencia en procesos quirúrgicos mediante el procesamiento de **imágenes hiper espectrales + clasificación** (DENEB)

Aplicaciones de los SEA (II)

Casos de uso desarrollados por ARCO

Plataformas autónomas

- Alargar tiempo de vida de la batería.
- Habilita la utilización de fuentes de energía renovables.

Foto-trampeo remoto (PLATINO)

Estimación de **poblaciones de insectos** en tiempo real (PLATINO)

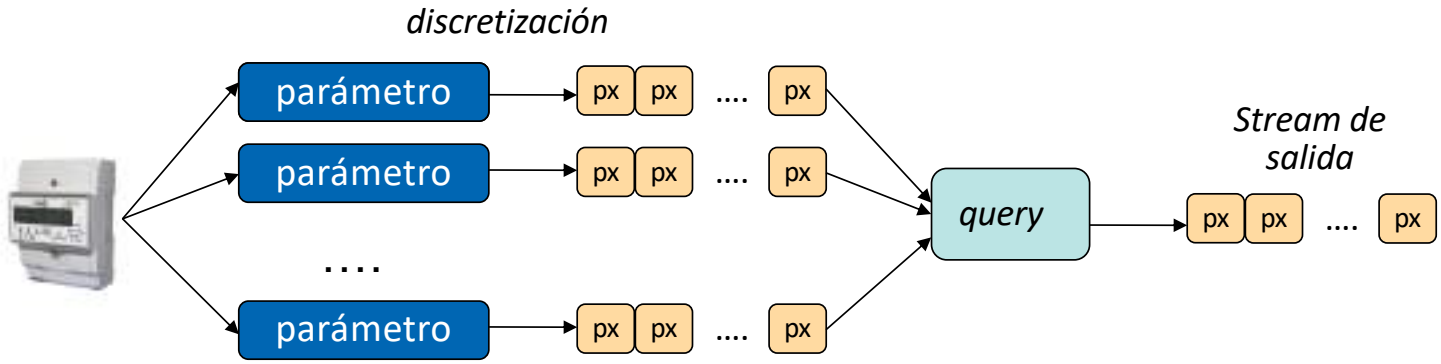
Nodo IoT con gestión energética inteligente (*energy harvesting*): algoritmos de **scheduling** para aplicaciones **neutrales en energía neutral** en **agricultura inteligente** (PLATINO)

Procesamiento - Algoritmos

Big Data e IA

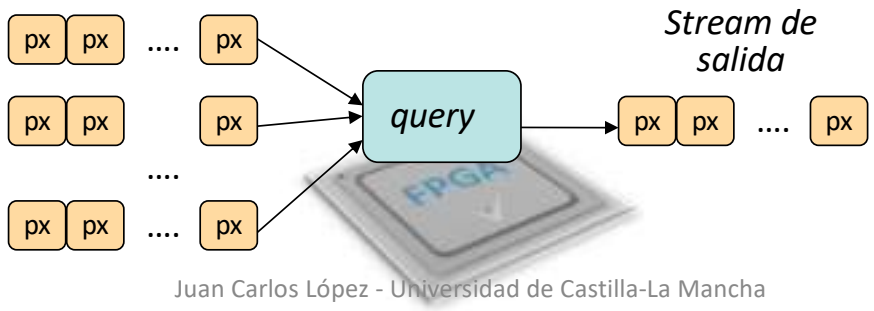
Aceleración Hw para computación Big Data

- Contexto
  - Datos procedentes de contadores eléctricos de 200M de usuarios
  - Múltiples parámetros medidos
    - Parámetros discretizados y almacenados en ficheros independientes
  - Lenguaje para la extracción de información basado en operaciones de conjuntos: unión, intersección, ...



Aceleración Hw para computación Big Data

- Basada en FPGAs
  - Ideal porque se trata de procesamiento en *streaming*
    - Construcción del operador de consulta a la medida
  - Tratamiento de múltiples *streams* simultáneamente
  - Compresión/descompresión de los ficheros sobre la marcha



Aceleración Hw para computación Big Data

- Procesamiento local para decisiones de gestión activa de la demanda

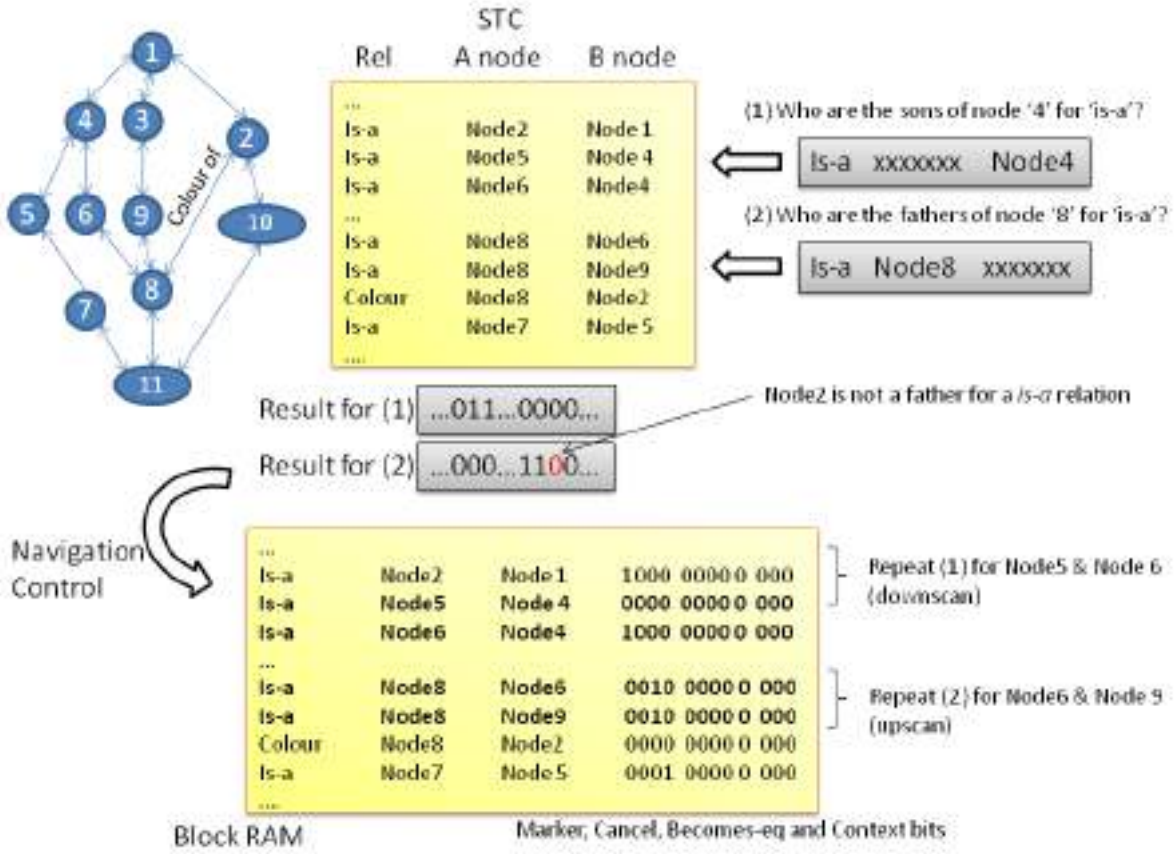
- En una etapa posterior:
  - Generación dinámica del operador a partir del lenguaje de interrogación:
    - Gracias a la tecnología de reconfiguración dinámica de las FPGAs
  - Despliegue sobre grid de FPGAs

Inferencia local: búsqueda y clasificación

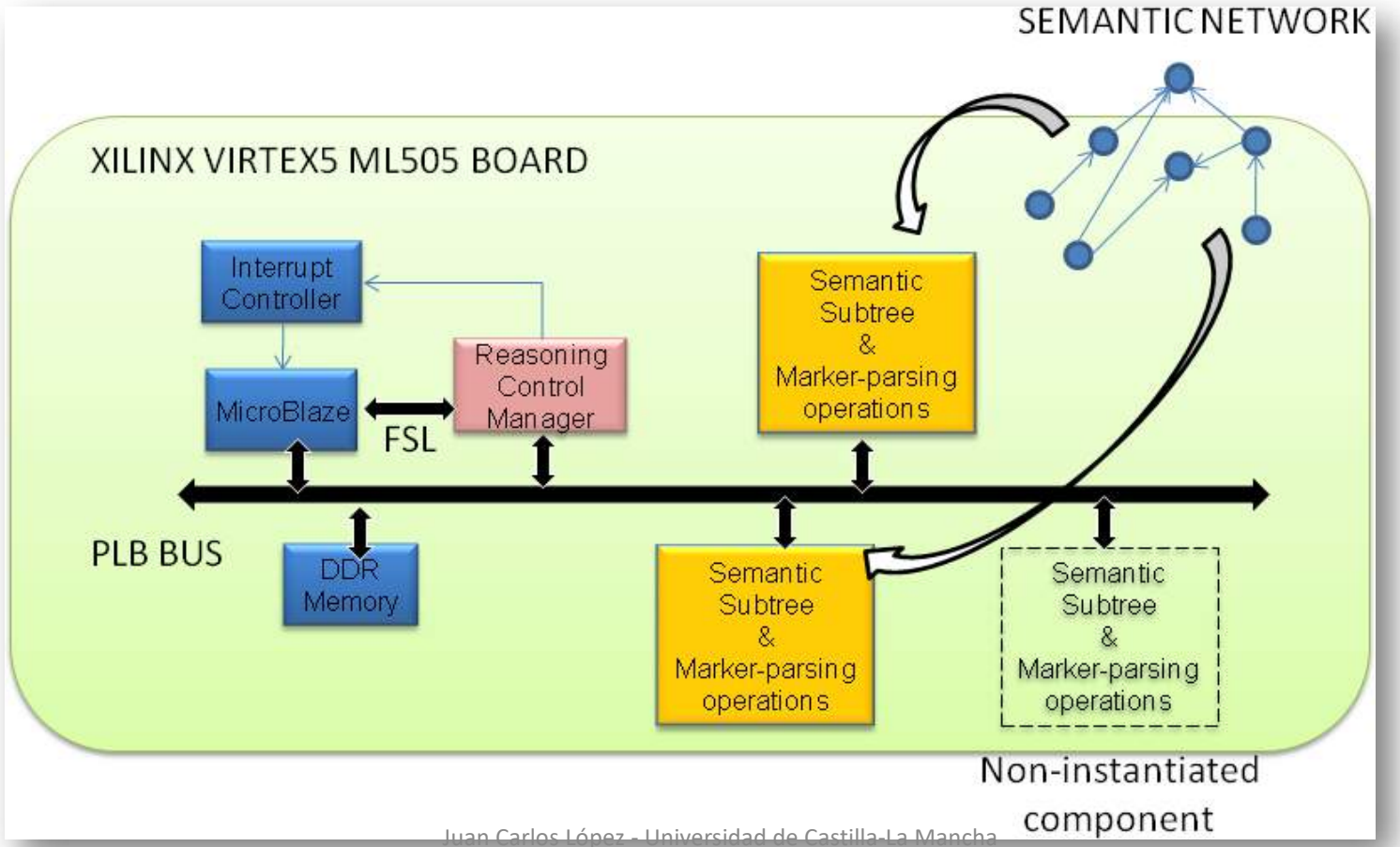
"Well, this is a dead end street. I have to brake 'till stop"

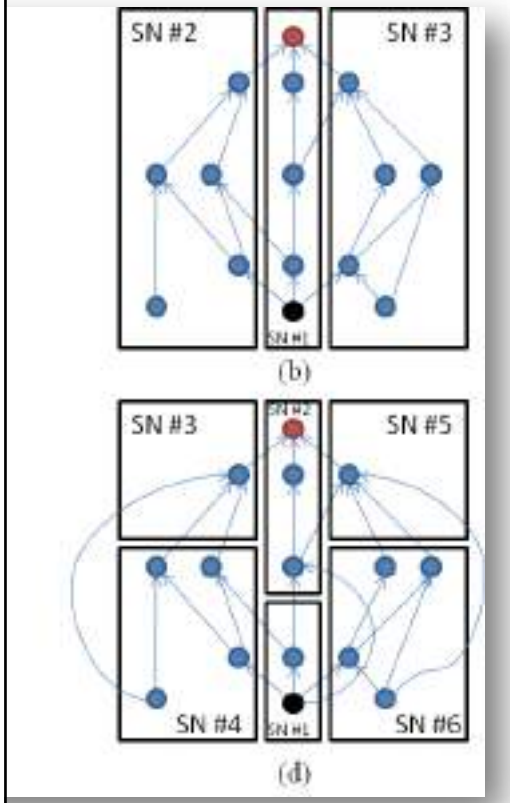
Inferencia local: búsqueda y clasificación

Razonamiento basado en sentido común básico



Scone System-on-Chip





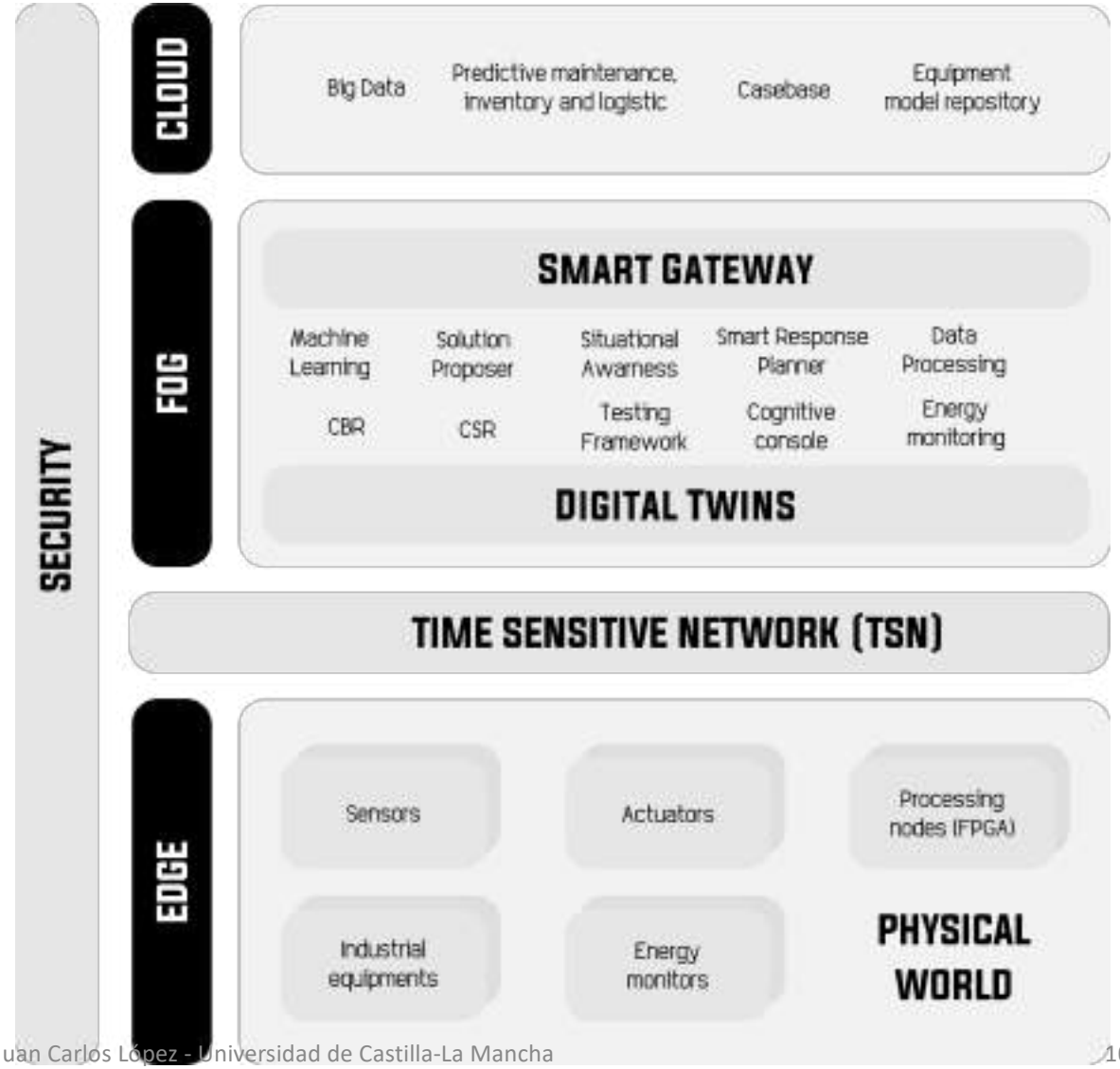
Test	Dell XPS 8300 (Intel i7-2600, 8GB DDR3,64 bit GNU-Linux)	HW Scone
Downscan the semantic network tree (1M elements)	4.5 sec	0.77 sec
Check the type of a given individual	0.23 msec	0.04 msec
Mark & intersect 2 sets with 10K members, one winner	20.71 msec	2.88 msec
Mark & intersect 3 sets with 10K members, one winner	36.9 msec	5.59 msec

ESCENARIOS REALES

Industria 4.0

Mantenimiento predictivo

Industria 4.0:
Mantenimiento predictivo



¡Muchas gracias!

Juan Carlos López

@jclopez2

juancarlos.lopez@uclm.es

http://arcoresearch.com