Optimización de escritorios virtuales

Sobre el software

Alberto Larraz Dalmases

Josep Maria Viñolas Auquer

IsardVDI

INSTITUT ESCOLA DEL TREBALL DE BARCELONA

.XTEC XarxaTelemàtica EducativadeCatalunya

Presentación

IsardVDI

- Instituto de Formación Profesional
  - 3500 alumnos
  - 240 profesores
  - 1200 ordenadores

- Diversidad de famílias profesionales y instalaciones.
- Dificultad para dedicar recursos a proyectos como este.
- Apuesta por software libre

Institut Escola del Treball de Barcelona

.XTEC XarxaTelemàtica EducativadeCatalunya

BACKGROUND

IsardVDI

- Escritorios tradicionales
  - 2002: SAMBA + openLDAP
  - 2010: iPXE Sistema de clonado centralizado. Desarrollo propio.
    - System Rescue CD + scripts python + BBDD
- Escritorios virtuales
  - **2013**: primeras pruebas y descartamos oVirt y GlusterFS
  - **2014-2016**: desarrollos propios con python y libvirt

SOLUCIÓN IsardVDI actual

KVM libvirt python

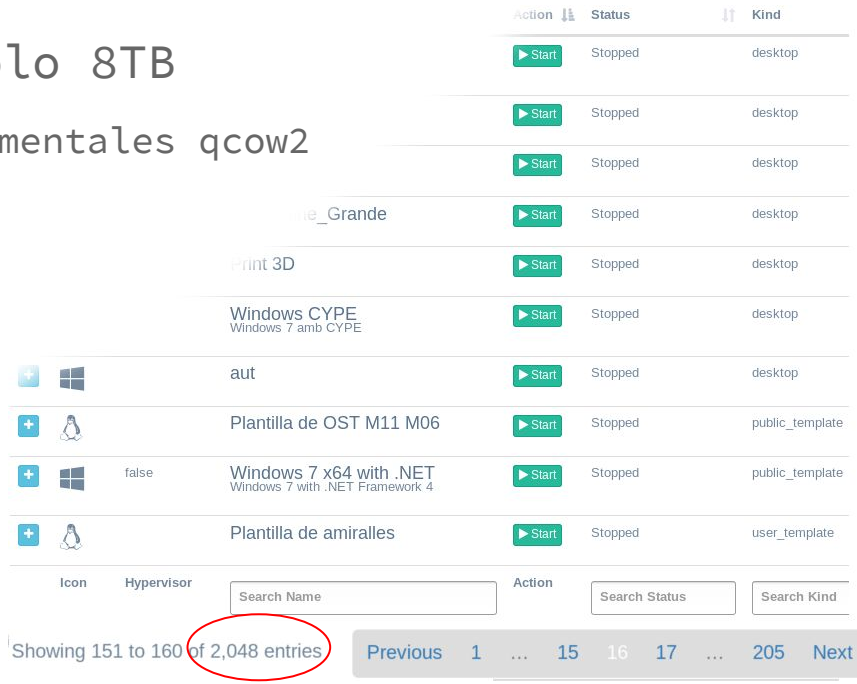
RethinkDB Flask Websockets

LOS NÚMEROS ACTUALES

IsardVDI

- **+1000 usuarios** usan IsardVDI en la escuela
- **+2000 escritorios virtuales** en 2 años (profes+alumnos)
- **+120 plantillas** creadas por el profesorado
- **+100 escritorios virtuales** simultáneos
- **+2100 discos** almacenados en sólo 8TB
  - Grácias al sistema de discos incrementales qcow2

- **6** hypervisores
- **46** Núcleos CPU / **92** Threads
- **576** GB memória RAM
- **16** TB en **2** NAS en cluster



OBJETIVOS

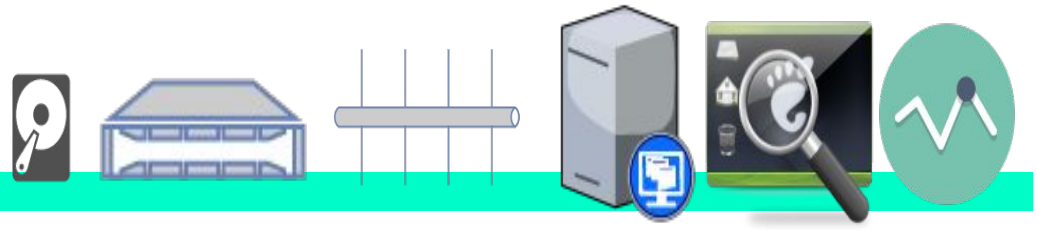
IsardVDI

- VDI ORIENTADO A:
  - Profesores y alumnos **autónomos** con sus escritorios
  - **Transición sencilla** a VDI en el aula
  - **Reducción de costes**

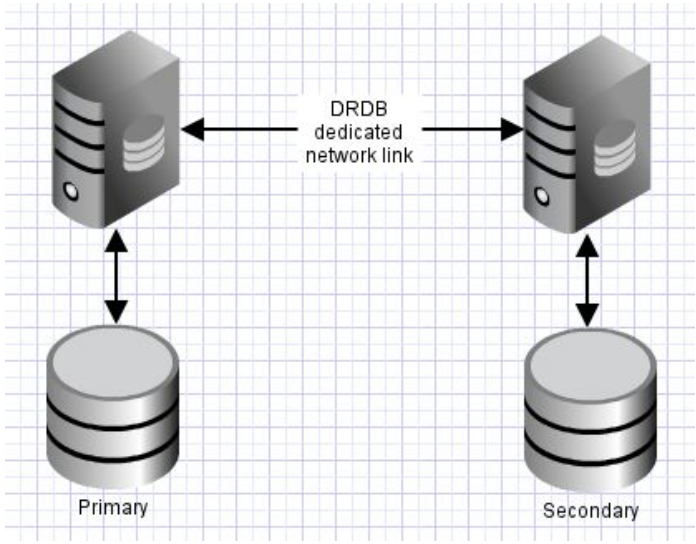
- OBJETIVOS TÉCNICOS PARA PROYECTO ISARD:
  - **Sistema de paths:** Independencia del almacenaje
  - **Pools de hypervisores:** Gestión de diferentes entornos
  - **Control centralizado libvirt+ssh:** Evitar software (daemons)
  - **Rápido en arranque y creación** de plantillas y escritorios
  - **Discos qcow incrementales:** Reducido espacio de almacenaje
  - **Protocolo spice:** Conexión transparente de dispositivos usb


Optimización

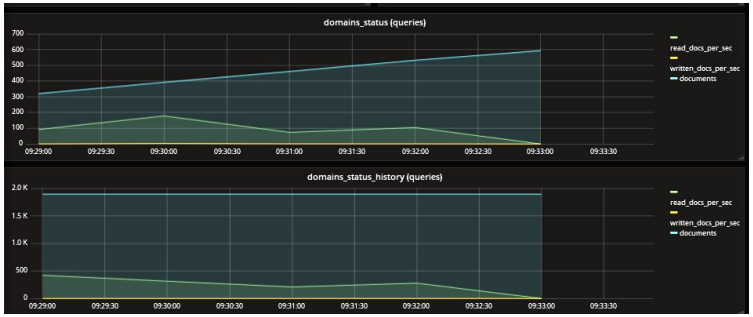
CAPAS DE OPTIMIZACIÓN



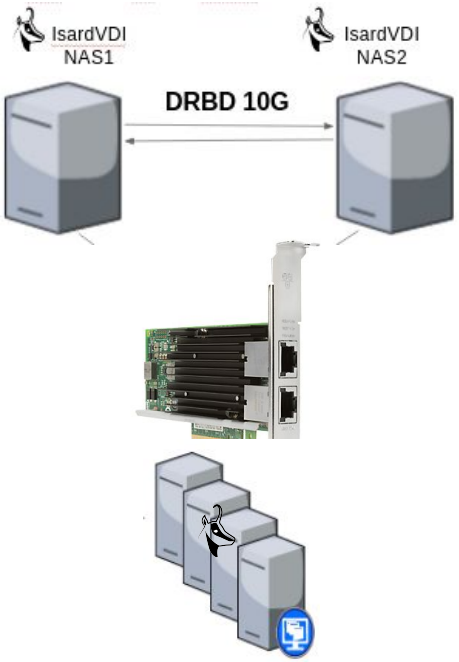
Ajustando las diferentes capas:



- Absorber picos de io en disco
- Redundancia y fiabilidad
- Red mínima 10G cobre
- Hypervisores DIY
- Lag y freeze en visores de cliente
- Prueba y error con monitorización constante



CAPAS DE OPTIMIZACIÓN



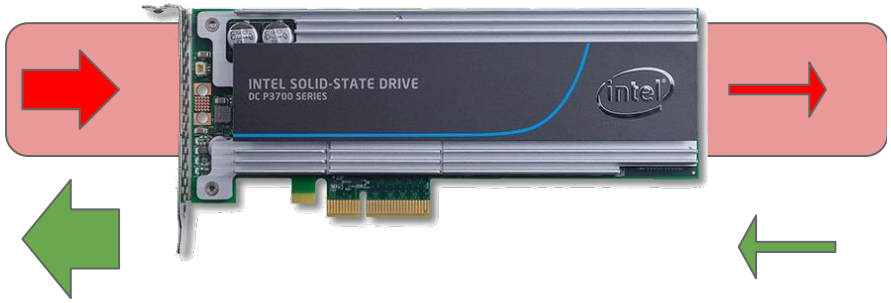
IsardVDI

- DISCOS:
  - CACHES
  - PARTICIONES y QCOWS
  - TESTS
- NAS
  - Balanceo de discos IsardVDI
- RED
  - Mapeo de IRQ
  - Stack TCP/IP
- HYPERVISORES
- VISORES
- MONITORIZACIÓN

CACHE I

rnd wr **160K** iops
seq wr **2GB/s**

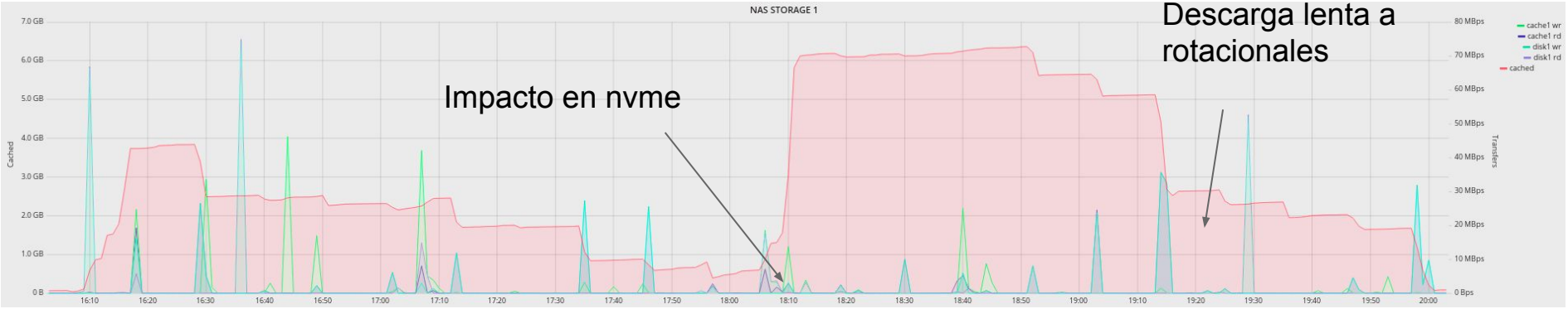
NVME / M2



ROTACIONAL

rnd **300K** iops
seq **5GB/s**

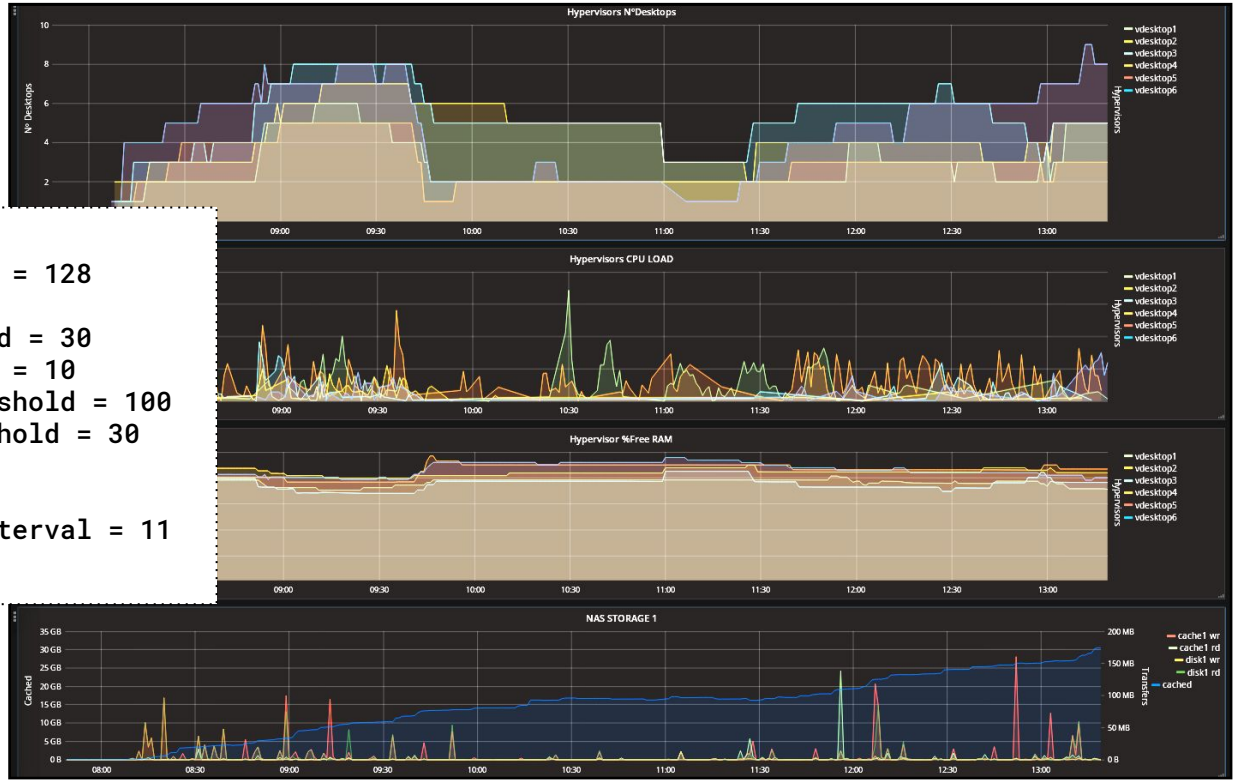
seq rd **160** iops
seq rd **200MB/s**



CACHE II

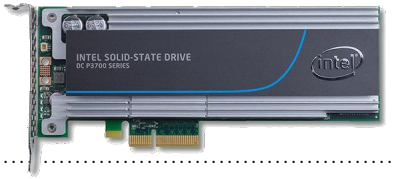
- **EnhanceIO**
  - Posibilidad de tunear comportamiento via **sysctl**
  - Permite monitorizar todos los parámetros de comportamiento, útil en optimización

```
[root@nas1]# sysctl -a | grep enhanceio
dev.enhanceio.data1.autoclean_threshold = 128
dev.enhanceio.data1.control = 0
dev.enhanceio.data1.dirty_high_threshold = 30
dev.enhanceio.data1.dirty_low_threshold = 10
dev.enhanceio.data1.dirty_set_high_threshold = 100
dev.enhanceio.data1.dirty_set_low_threshold = 30
dev.enhanceio.data1.do_clean = 0
dev.enhanceio.data1.mem_limit_pct = 75
dev.enhanceio.data1.time_based_clean_interval = 11
dev.enhanceio.data1.zero_stats = 0
```



DISCOS

- FIO: Tests especificaciones

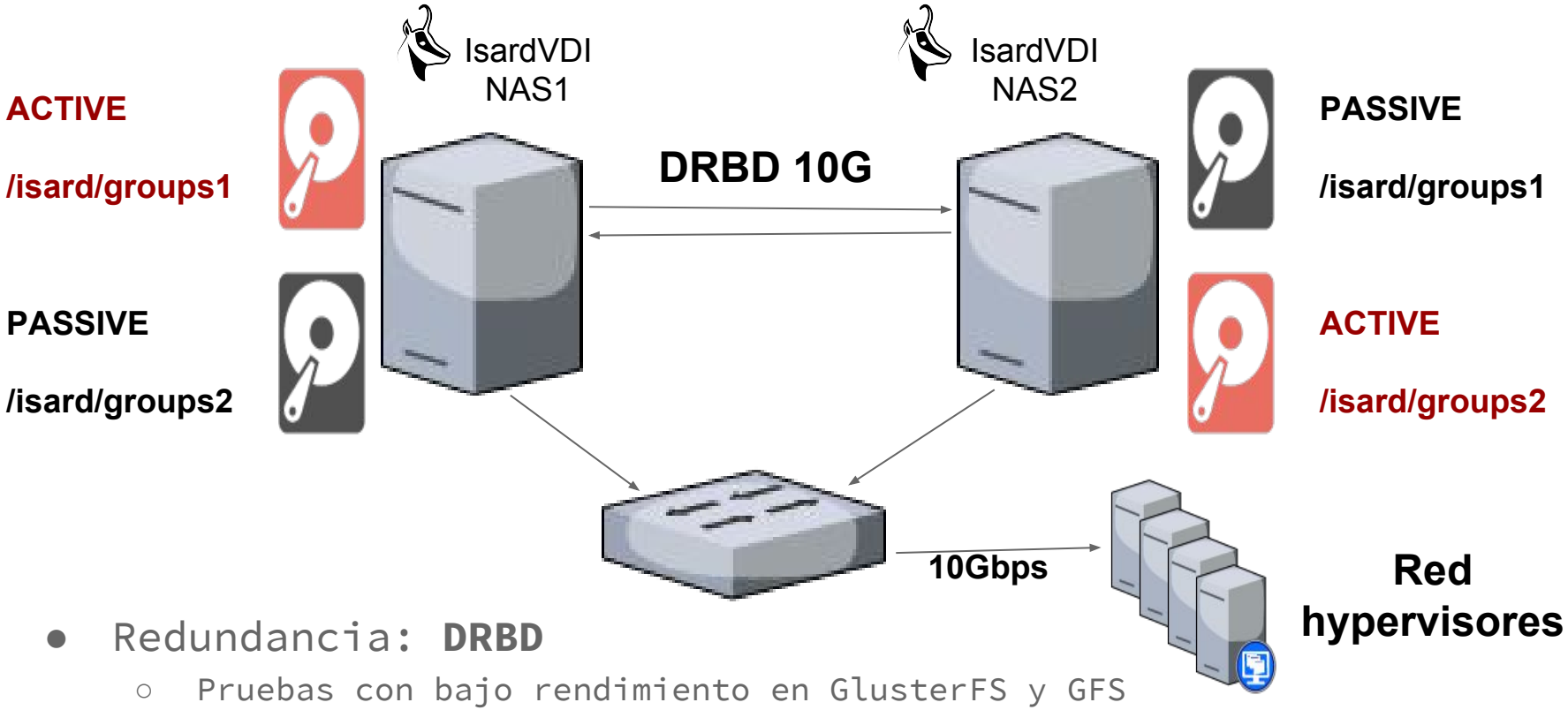



```
# MODEL                SIZE  SEQREAD_128KB  SEQWR_128KB  RANDREAD_4K  RANDWR_4K
# Intel 750 series     400G  2.200MB/S              900MB/S              430.000
230.000
# PCIe x4 # Intel i5-6500 8GT/S. 32GB RAM. Gygabyte GA-Z170-Gaming7-EU.
```

```
READ 128K (MAX BW)
   READ: io=40960MB, aggrb=2177.2MB/s, minb=2177.2MB/s, maxb=2177.2MB/s, mint=18814msec,
maxt=18814msec
RANDOM READ 4K (MAX IOPS)
   READ: io=29406MB, aggrb=1470.3MB/s, minb=1470.3MB/s, maxb=1470.3MB/s, mint=20001msec,
maxt=20001msec
WRITE 128K (MAX BW)
  WRITE: io=19572MB, aggrb=976.99MB/s, minb=976.99MB/s, maxb=976.99MB/s, mint=20033msec,
maxt=20033msec
RANDOM WRITE 4K (MAX IOPS)
  WRITE: io=9133.5MB, aggrb=467610KB/s, minb=467610KB/s, maxb=467610KB/s, mint=20001msec,
maxt=20001msec
RANDOM READ/WRITE 64K (SIMULATED QCOW2)
   READ: io=13225MB, aggrb=676765KB/s, minb=676765KB/s, maxb=676765KB/s, mint=20010msec,
maxt=20010msec
  WRITE: io=13185MB, aggrb=674734KB/s, minb=674734KB/s, maxb=674734KB/s, mint=20010msec,
maxt=20010msec
```

NAS

- IsardVDI balancea creación y provisión de discos
  - Reparto de red y de io de disco



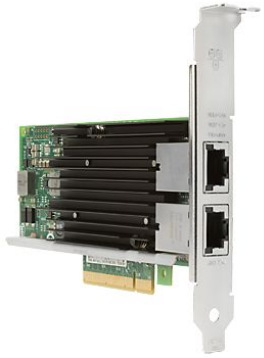
- Redundancia: **DRBD**
  - Pruebas con bajo rendimiento en GlusterFS y GFS
  - Demasiada inversión para montar CEPHS
- Fiabilidad: **Pacemaker**

OPTIMIZACIÓN: RED I

- Cuellos de botella:

>2GBps = >16Gbps

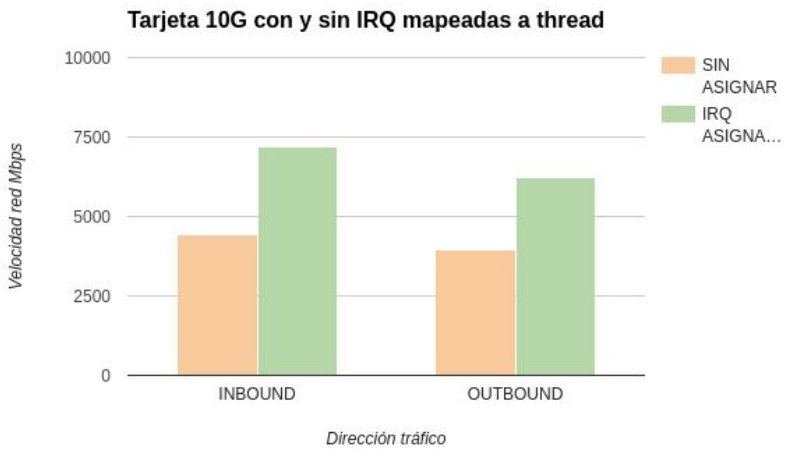
Mapeo de IRQ



2x10G

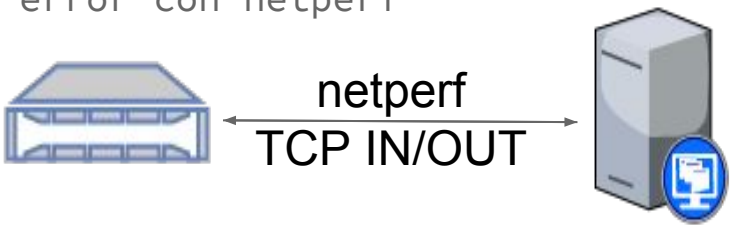
IRQ SIN MAPEAR: Kernel, lento
IRQ MAPEADA: Hardware, rápido

PCI:
- v. 3.x (8 GT/s):
  - 985 MB/s (×1)
  - 15.75 GB/s (×16)
- v. 4.0 (16 GT/s):
  - 1.969 GB/s (×1)
  - 31.51 GB/s (×16)



OPTIMIZACIÓN: RED II

- STACK TCP para NFS/10G
  - Prueba y error con netperf



Parámetros sysctl stack tcp/ip

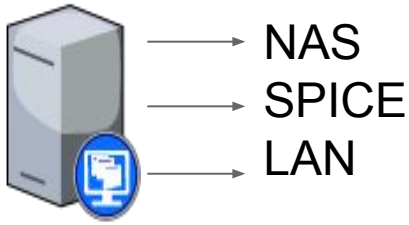
```
net.ipv4.tcp_sack = 0
net.ipv4.tcp_timestamps = 0

net.ipv4.tcp_rmem = 10000000 10000000 10000000
net.ipv4.tcp_wmem = 10000000 10000000 10000000
net.ipv4.tcp_mem = 10000000 10000000 10000000

net.core.rmem_max = 524287
net.core.wmem_max = 524287
net.core.rmem_default = 524287
net.core.wmem_default = 524287
net.core.optmem_max = 524287
net.core.netdev_max_backlog = 300000
```

NETPERF: Simultáneo in/out TCP 256k 10 segundos

```
for i in 1
 do
  netperf -H 10.1.2.31 -t TCP_STREAM -B "outbound" \
    -i 10 -P 0 -v 0 -- -s 256K -S 256K &
  netperf -H 10.1.2.31 -t TCP_MAERTS -B "inbound"  \
    -i 10 -P 0 -v 0 -- -s 256K -S 256K &
 done
```

HYPERVISORES

- CPU/THREADS vs CLOCK SPEED
  - Pruebas con i7: Gigabyte gaming
  - Pruebas con Xeon dual socket:
    - Intel Server & Supermicro
- COSTE/REDIMIENTO
  - Según mercado hay que escojer:
    - Densidad CPU y RAM = Elevado coste
    - Velocidad CPU y RAM limitada = Coste contenido
- SEPARAR TRÁFICOS DE RED
  - NAS: Acceso a disco >= 10G
  - SPICE: Visores cliente >= 10G
  - LAN: Acceso Internet vms >= Nx1G



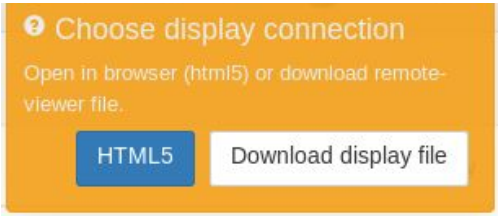
IsardVDI Hypervisor

OPTIMIZACIÓN: VISORES

- **SPICE**
  - Hay que adaptarlo a la app específica que correrá el escritorio
  - Tráfico VBR (2Mbps-120Mbps)

```
<graphics type='spice' port='...>
      ….
      <image compression='auto_glz'/>
      <jpeg compression='auto'/>
      <zlib compression='auto'/>
      <playback compression='on'/>
      <streaming mode='filter'/>
</graphics>
```

- **HTML5**
  - Requiere ProxySocket. Puede ser cuello de botella y “lag”.
  - Posibles situaciones:
    - **Hypervisor**
    - **Servidor dedicado**

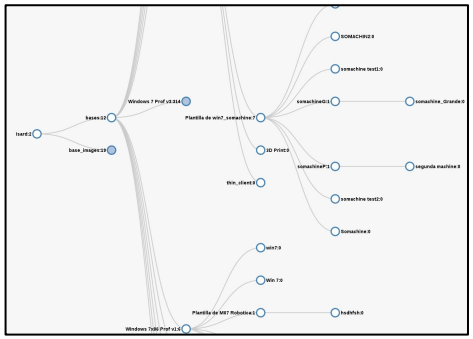


- **RDP**
  - Buen comportamiento bajo escritorios Windows

OPTIMIZACIONES: Priorizar (802.1p) y separar (802.1q) tráfico de vídeo. Clientes 1G. Servidores y troncales 10G o bonding.

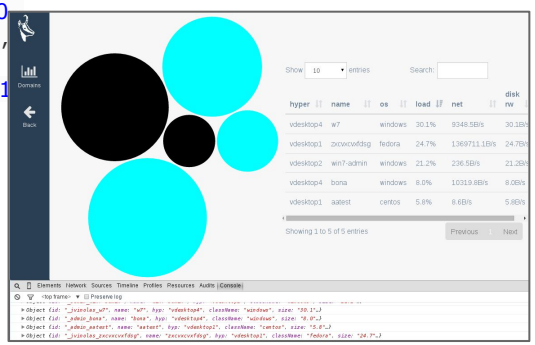
MONITORIZACIÓN I

Desktops

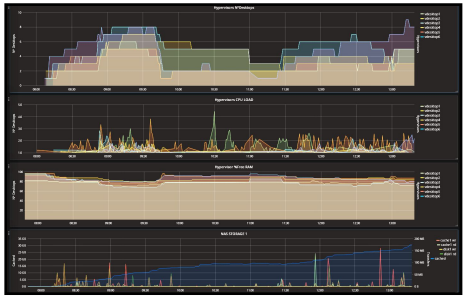



```
"raw_stats": {
        "balloon.current": 3768320 ,
        "balloon.last-update": 1496829053 ,
        "balloon.maximum": 3768320 ,
        "balloon.rss": 3912476 ,
        "balloon.swap_in": 0 ,
        "block.0.allocation": 39517487104 ,
        "block.0.capacity": 48318382080 ,
        "block.0.fl.reqs": 58879 ,
        "block.0.fl.times": 32000232152 ,
        "block.0.name": "vda" ,
        "block.0.path":
        "/vdisks_local/f23dev-adria.qcow2" ,
        "block.0.physical": 39517495296 ,
        "block.0.rd.bytes": 2289736704 ,
        "block.0.rd.reqs": 76716 ,
        "block.0.rd.times": 270308759123 ,
        "block.0.wr.bytes": 4135097344 ,
        "block.0.wr.reqs": 219527 ,
        "block.0.wr.times": 110816053641 ,
        "block.1.allocation": 0 ,
        "block.1.fl.reqs": 0 ,
        "cpu.system": 308500000000 ,
        "cpu.time": 302127770015
        "cpu.user": 527280000000 ,
        "net.0.name": "vnet1" ,
        "net.0.rx.bytes": 153577521
        "net.0.rx.drop": 0 ,
        "net.0.rx.errs": 0 ,
        ...
```

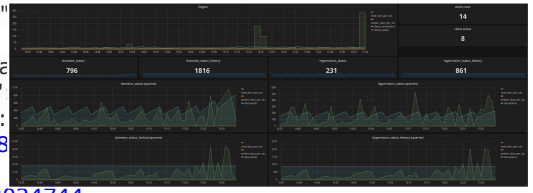

IsardVDI



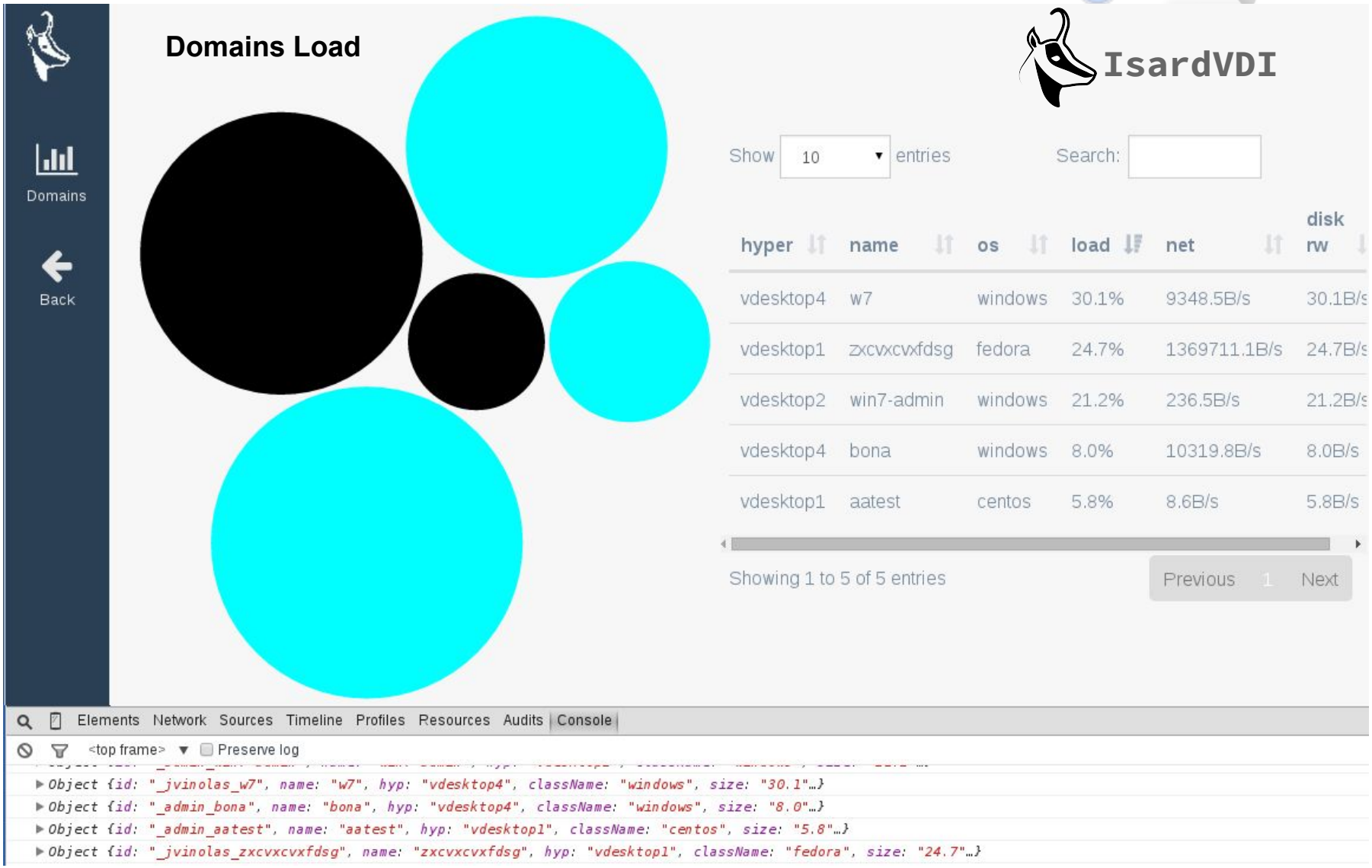
Hypervisors



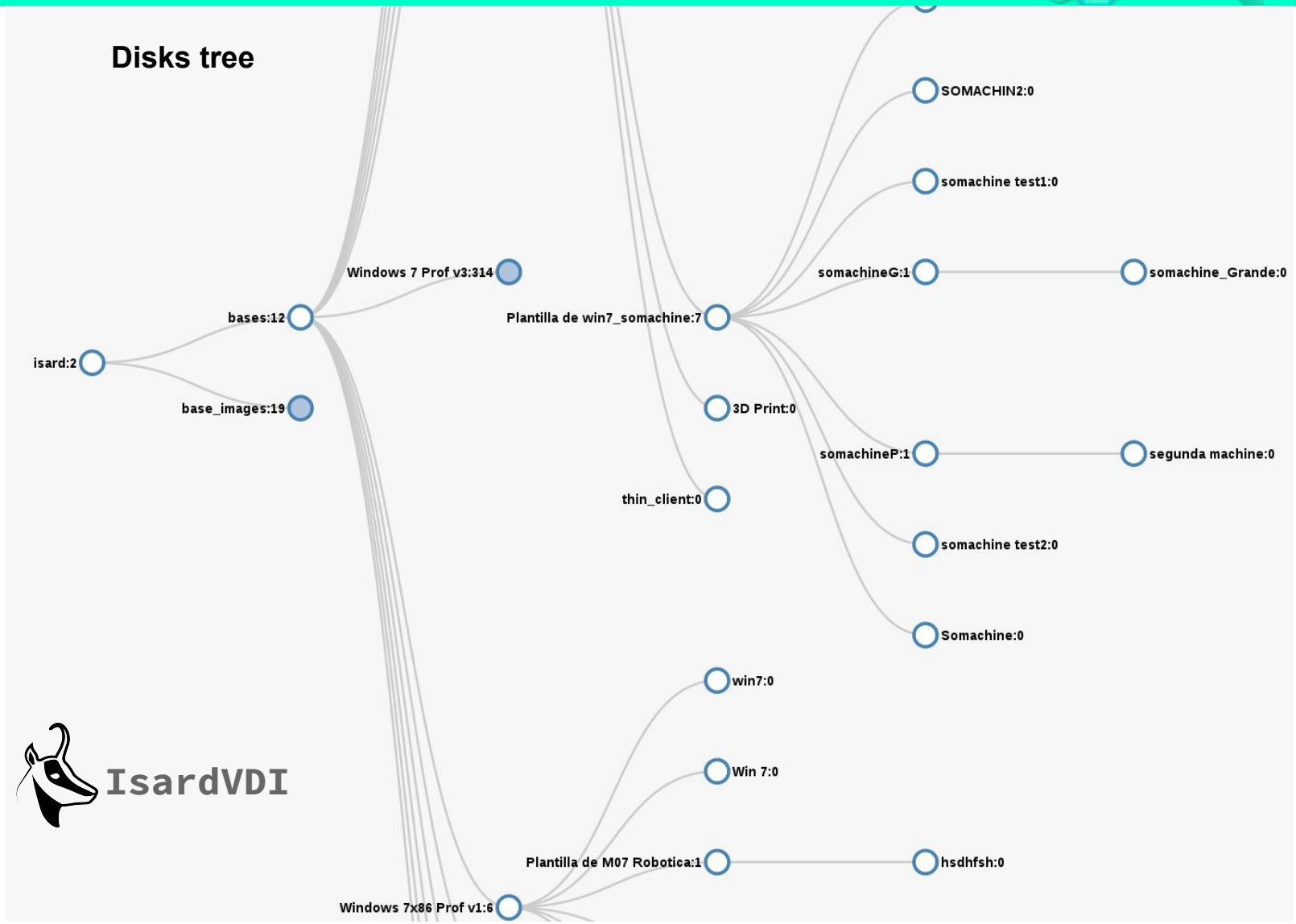

```
{
"connected": true ,
"cpu_percent": {
        "idle": 99.951 ,
        "iowait": 0 ,
        "kernel": 0.041 ,
        "used": 0.049 ,
        "user": 0.008
} ,
"delay_from_connect":
0.007673978805541992 ,
"delay_query_load":
0.0066356658935546875 ,
"domains": [ ],
"hostname": "vdesktop2.escoladeltreball.org"
,
"hyp_id": "vdesktop2" ,
"id":
"084d3988-1c93-451a-9fb3-0d6b06da3495" ,
"load": {
        "cpu_load": {
                "idle": 88119475310000000
                ,
                "iowait": 26875000000000 ,
                "kernel":
                1020701460000000 ,
                "user"
        } ,
        "free_ram_tota
        "percent_free"
        "ram_cached":
        "ram_free": 18
} ,
"when": 1496927518.2024744
}
```

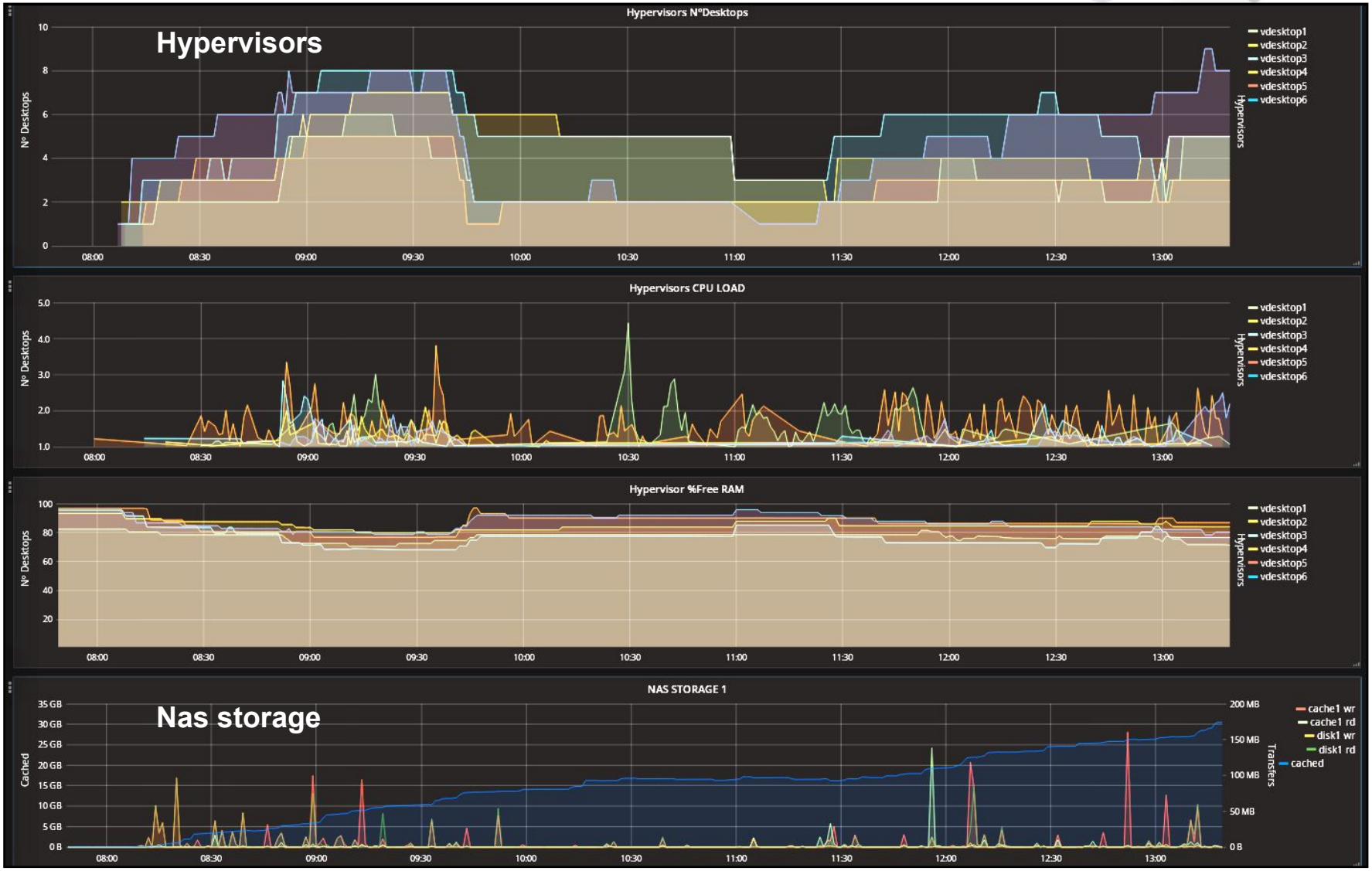
MONITORIZACIÓN II



MONITORIZACIÓN III



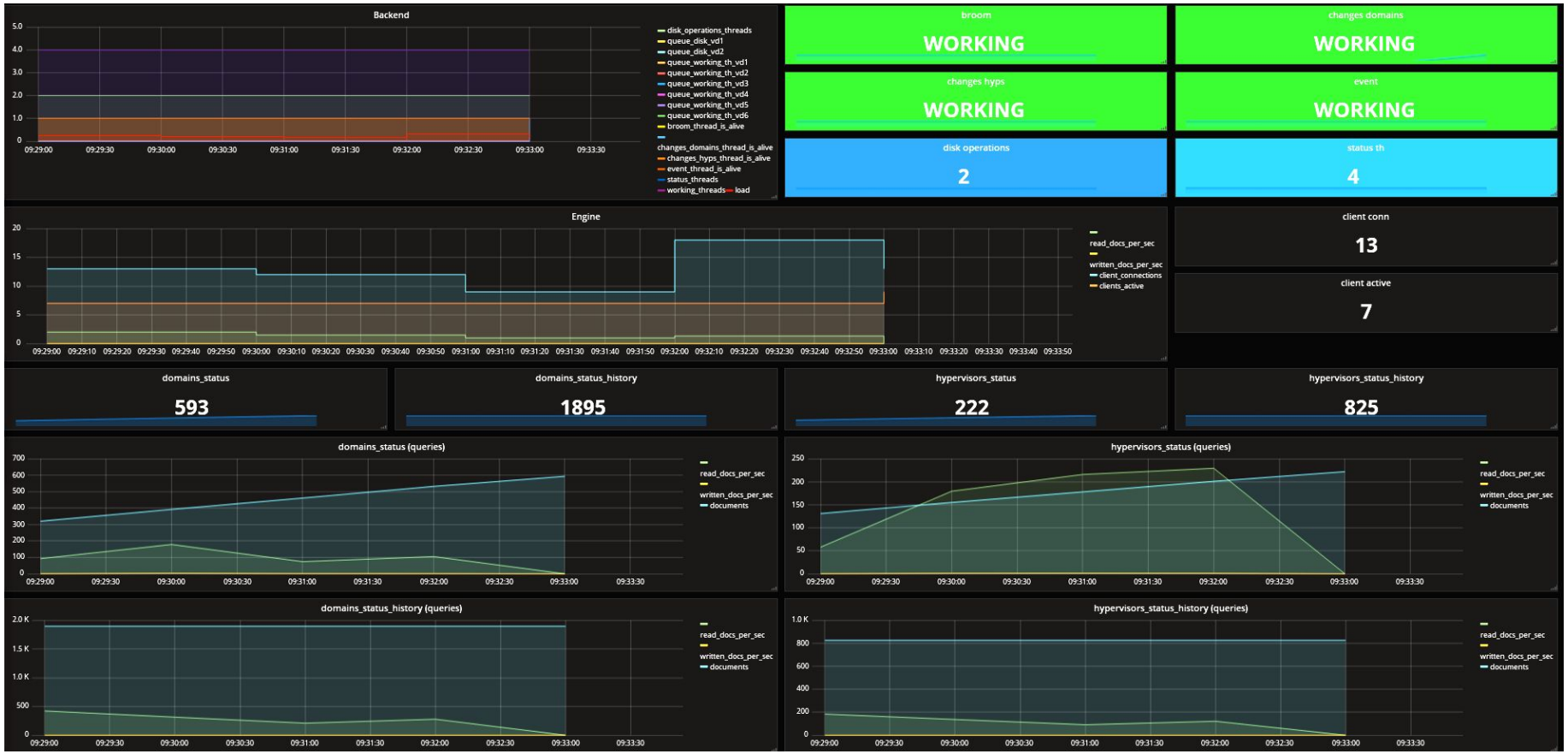
MONITORIZACIÓN IV



MONITORIZACIÓN V

IsardVDI

Database & Engine



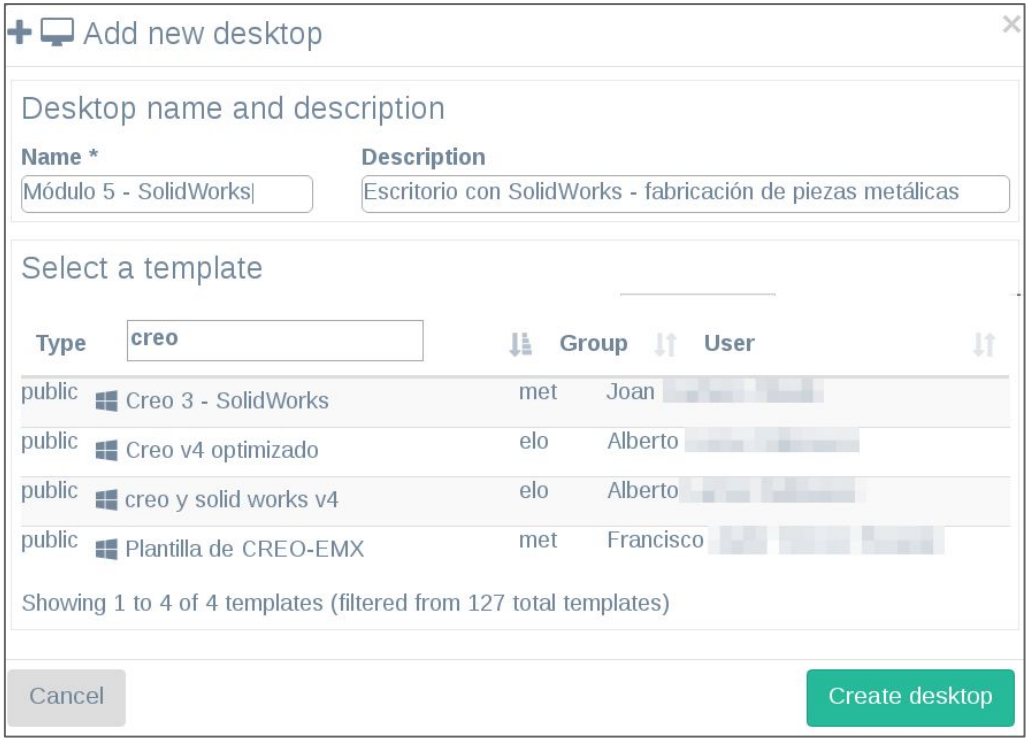
ISARD VDI

ORIENTADO A USO DOCENTE (I)

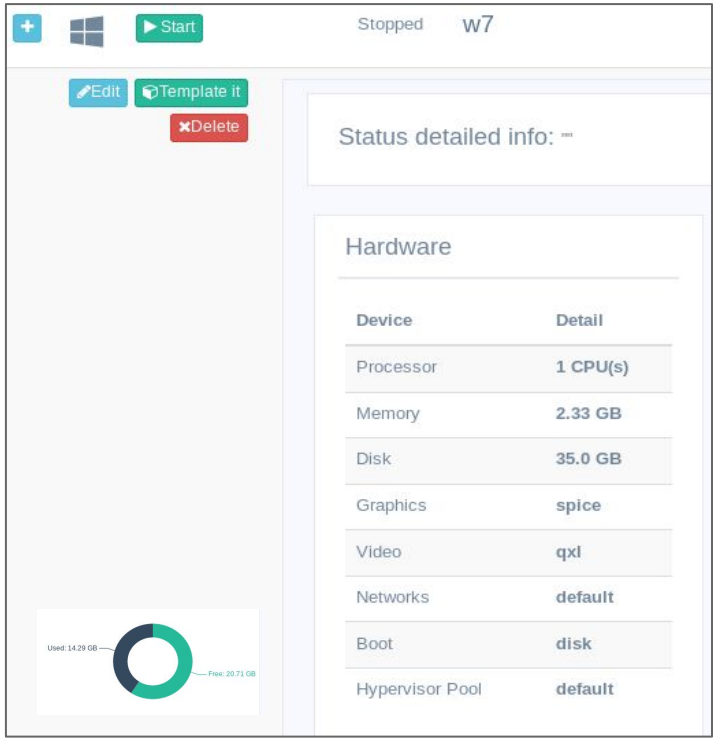
IsardVDI

- Profesores autónomos: pueden crear y compartir plantillas, orientado a despliegue inmediato en el aula

Creación de escritorio a partir de plantilla



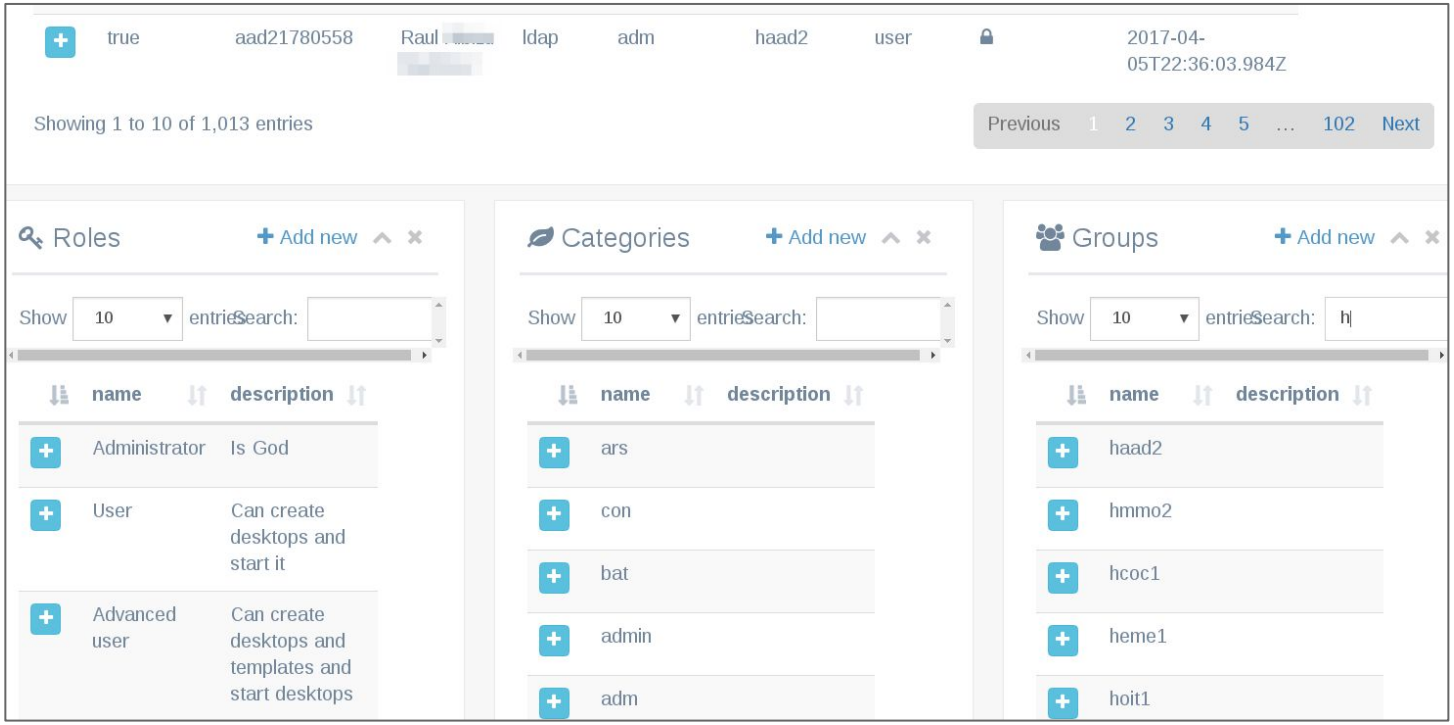
Generar plantilla de escritorio



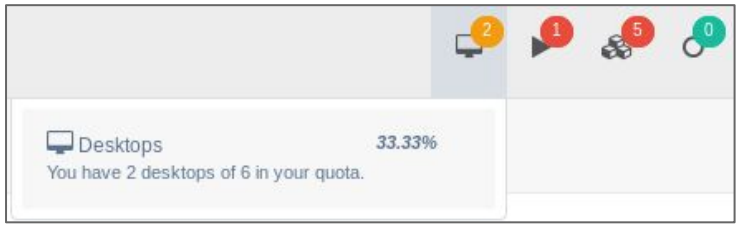
ORIENTADO A USO DOCENTE (II)

IsardVDI

- Usuarios por LDAP, clasificados en categorías / grupos



- Quotas por grupos para crear, arrancar



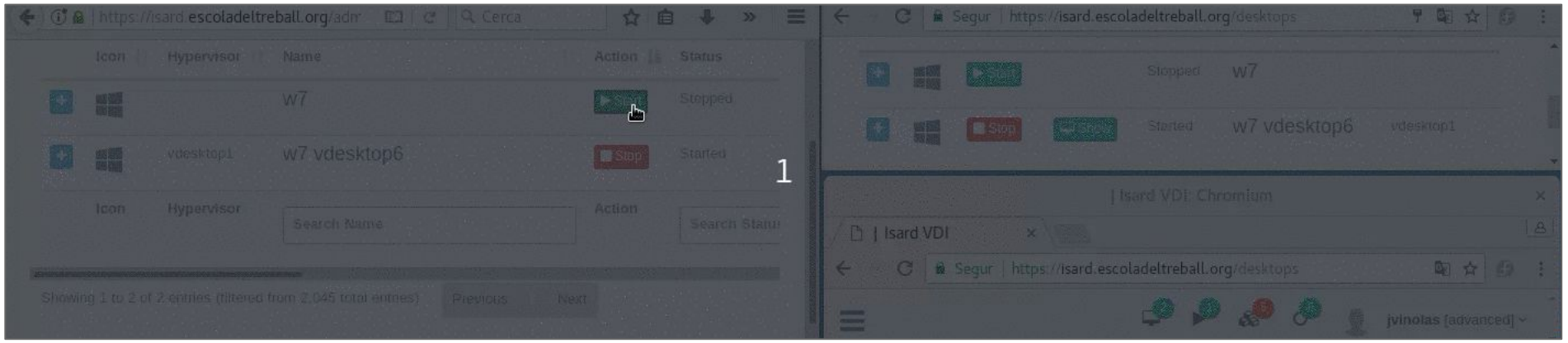
ORIENTADO A USO DOCENTE (III)

IsardVDI

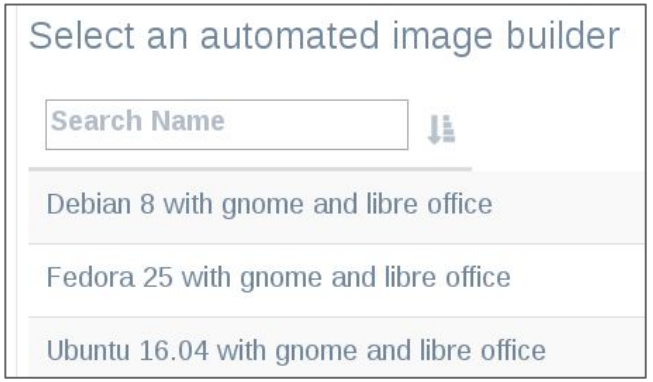
- Actualización en tiempo real de la UI (websockets)

Administrador IsardVDI

Usuario IsardVDI



- Potenciar el uso de distribuciones de linux actualizadas

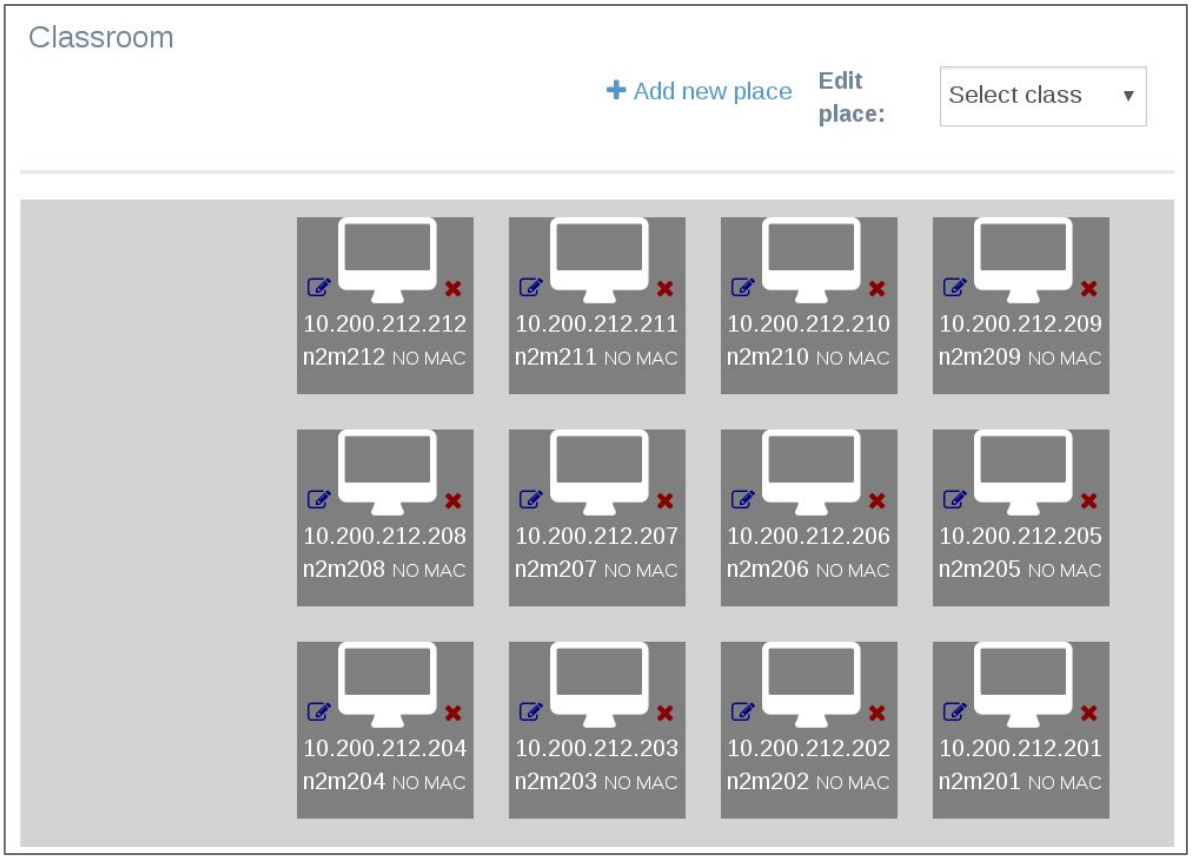


ORIENTADO A USO DOCENTE (VI)

*En desarrollo

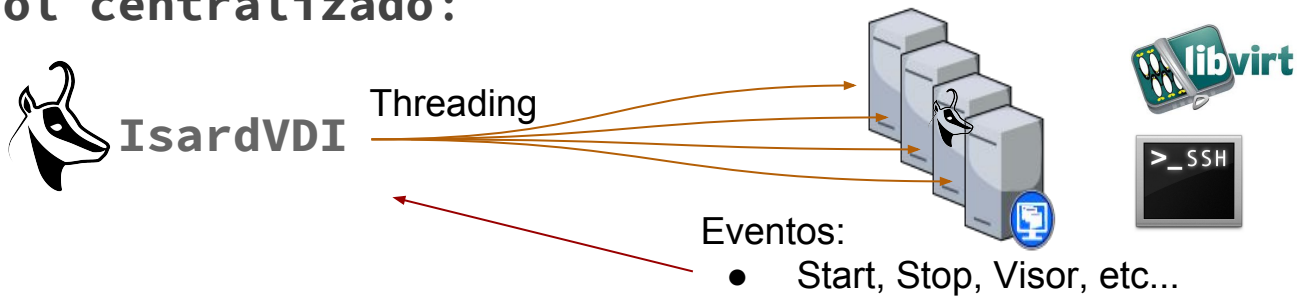
IsardVDI

- Administrar un aula real o virtual
  - Interfaz de usuario con eventos en tiempo real
- Crear y arrancar masivamente escritorios
  - Más control del aula por el profesor.



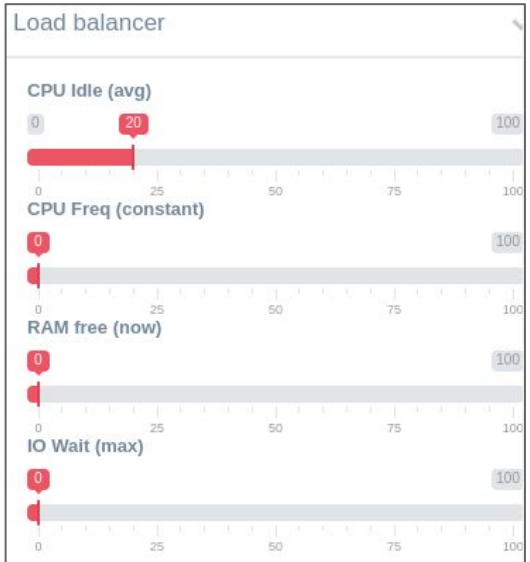
Engine propio

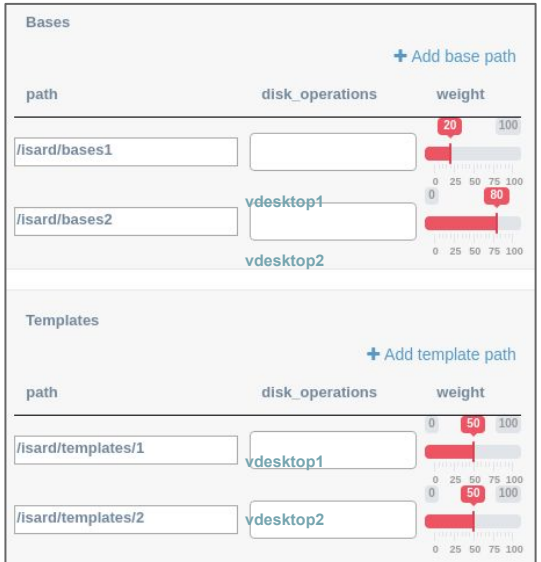
- Desarrollado pensando en el **uso en las aulas**
- **Control centralizado:**



- Podemos exprimir las posibilidades que da libvirt/qemu
- **Pools de hypervisores:** balanceo y orquestración con pesos

```
"weights": {
      "avg_cpu_idle": {
            "parameter": "cpu_idle" ,
            "type": "average" ,
            "weight": 50 } ,
      "randomize": {
            "parameter": "random" ,
            "type": "current" ,
            "weight": 10 } ,
      "vcpus_rcpus": {
            "parameter": "rate_vcpus_rcpus"
            ,
            "type": "current" ,
            "weight": 40 }
```

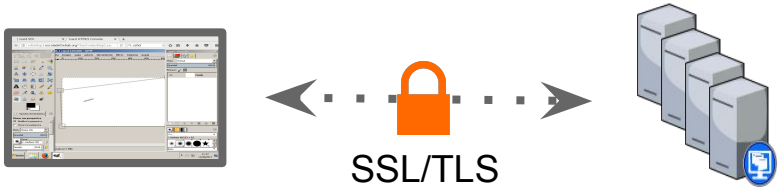




Clientes SPICE

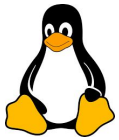
IsardVDI

- Tráfico cifrado TLS



- Reutilización PCs
  - PXE: Arranque por red

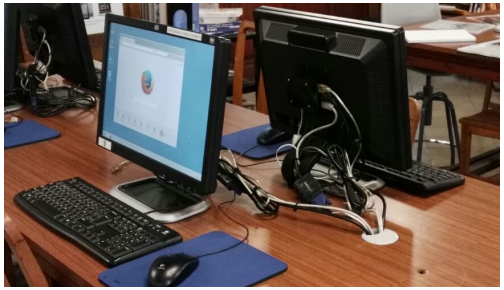
- Multiplataforma





- BYOD y Clientes ligeros
  - El alumno trae su dispositivo
  - Reducción costes
    - Eléctricos
    - Renovación

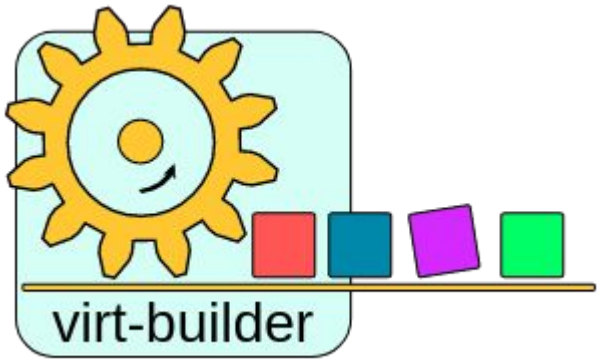


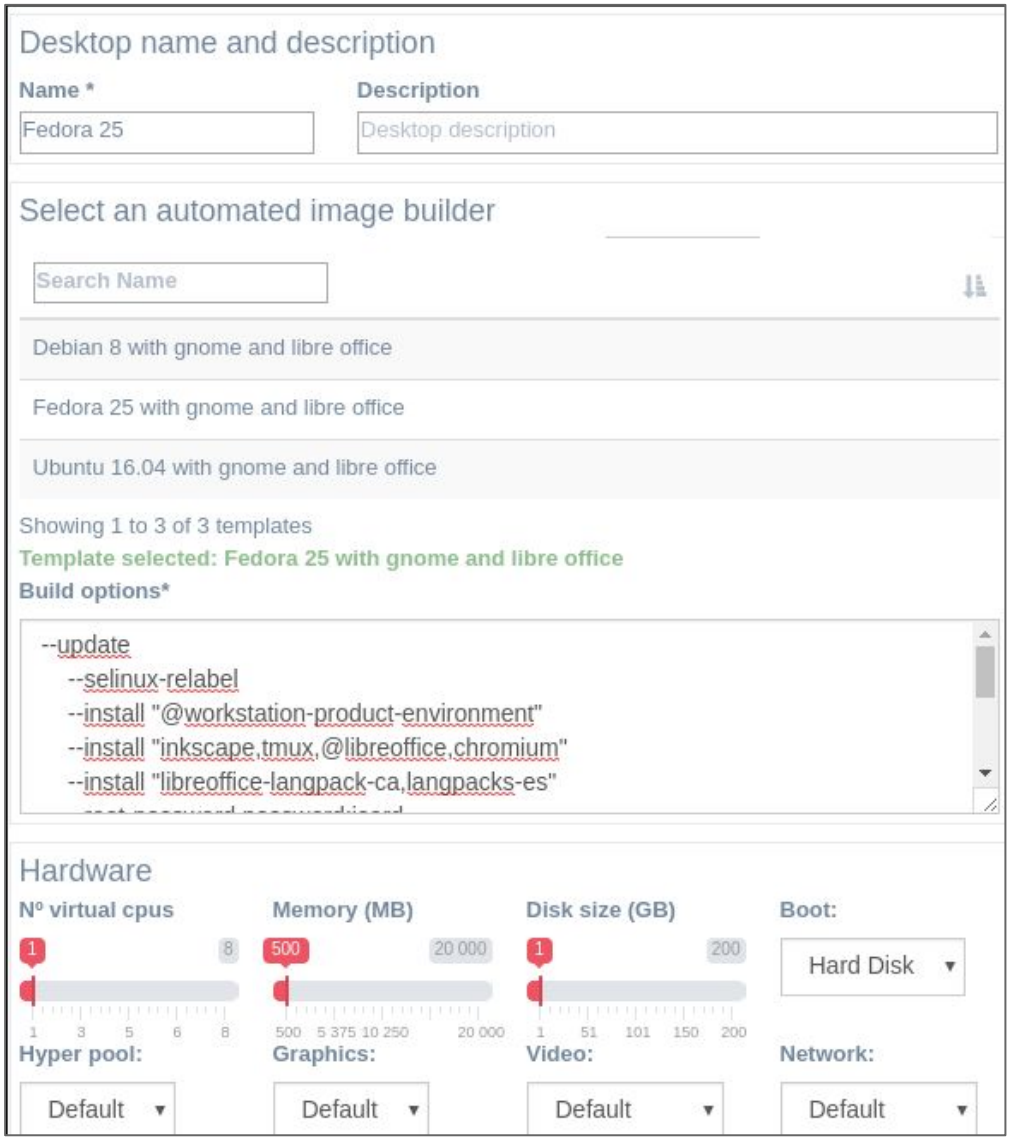


Auto Builder

IsardVDI

- Creación automatizada de escritorios
- Configurable



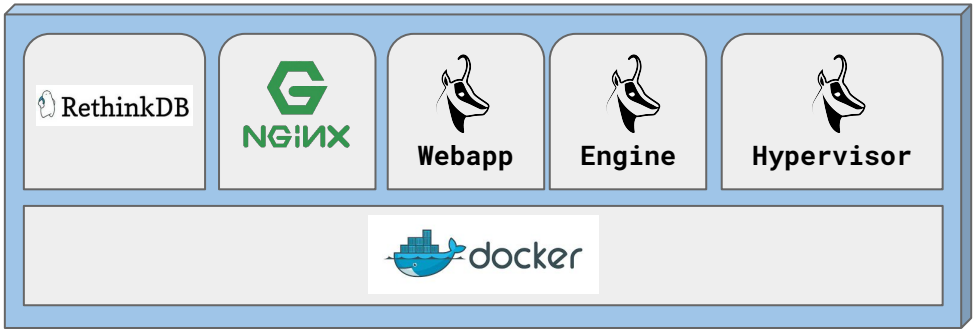


UN IsardVDI PARA CADA CASO

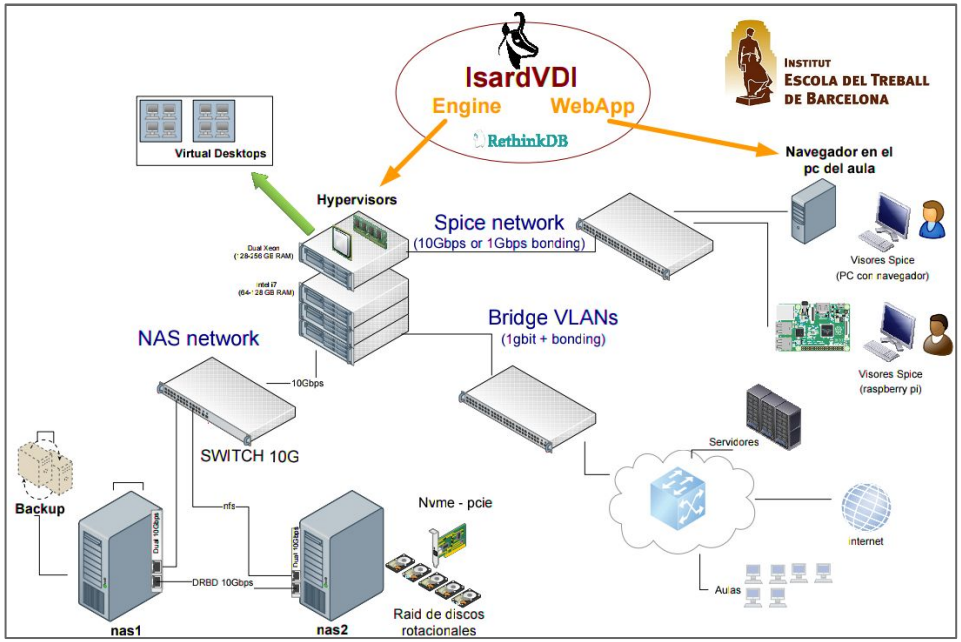
IsardVDI

Rápida implementación con Docker

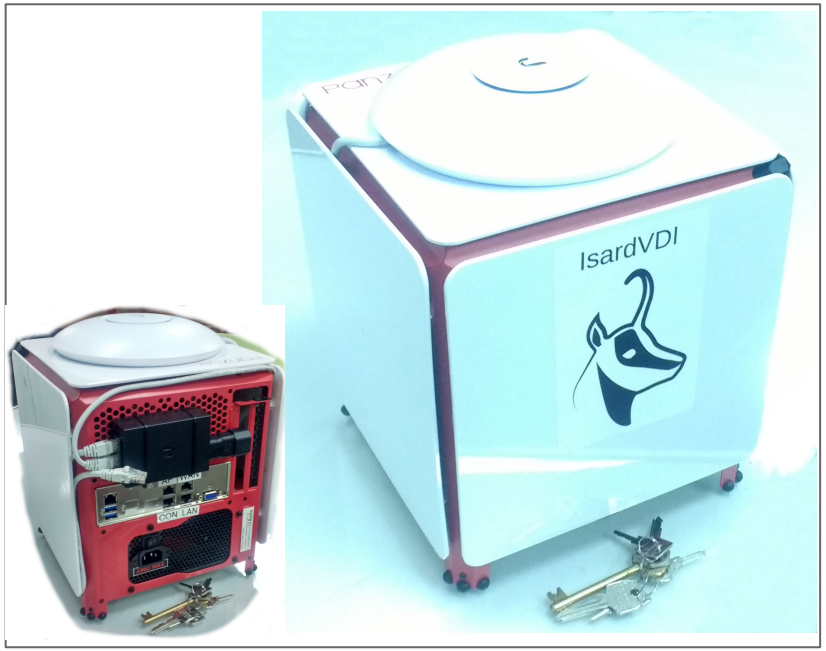
```
git clone https://github.com/isard-vdi/isard.git
cd isard & docker-compose up
```



Infraestructura distribuida



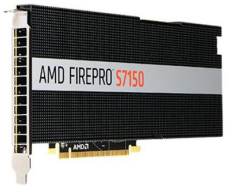
All-in-one



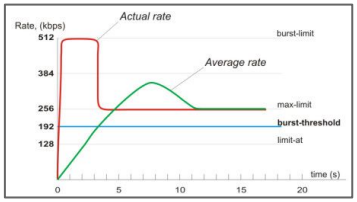
NUEVAS FUNCIONES EN PRUEBAS

IsardVDI

- Virtualización **GPU** (kernel >=4.11): VFIO-MxGPU (AMD)
  - **PCI passthrough** pools



- Limitación de **io por escritorio** en funcionamiento.



- **Medición del rendimiento** de una infraestructura de VDI
- Unidad virtual compartida entre escritorios
- **Live migration** de escritorios entre hypervisores
- Arranque y apagado de hypervisores dinámicamente

Objetivos próximos

IsardVDI

- Pasar del 50% al **99% de usuarios** en la Escola del Treball
- Conseguir recursos y crear sinergias
- Consolidar software y darle robustez
  - Necesitamos developers
- Implementar pilotos
- Mejorar difusión y documentación
- Probar nuevo hardware (hypervisores, nas, clientes)

Muchas gracias por su atención…
Alguna pregunta?

Alberto Larraz Dalmases
alarraz@escoladeltreball.org

Josep Maria Viñolas auquer
jvinolas@escoladeltreball.org

DEMO ONLINE!

WEB: **https://try.isardvdi.com**
VOUCHER CODE: **rediris2017isard**

PRUEBA EN TU LINUX:

```
git clone https://github.com/isard-vdi/isard.git
cd isard & docker-compose up
```

IsardVDI

INSTITUT ESCOLA DEL TREBALL DE BARCELONA

XTEC XarxaTelemàtica EducativadeCatalunya