

# **Xen and the art of virtualization (2003)**

Paul Barham , Boris Dragovic, Keir Fraser, Steven Hand,  
Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, Andrew Warfield

# Contenido

1. Introducción
2. La propuesta de Xen y sus características generales
3. Detalles del diseño de Xen
4. Evaluación
5. Trabajo Futuro
6. Crítica



# Introducción

The background features abstract, flowing shapes in shades of orange and red. On the left, there are overlapping orange waves. On the right, there are overlapping red waves. These shapes create a sense of movement and depth, framing the central text.

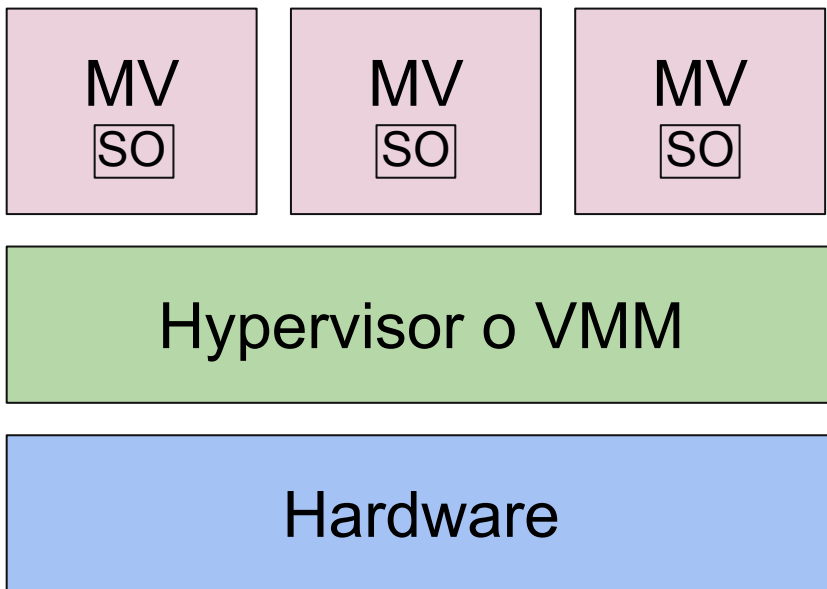
# Máquinas virtuales (MV)

- Las MV deben estar aisladas
- Deben dar soporte a distintos SO
- La sobrecarga por virtualización debe ser pequeña



# ¿Qué es Xen?

- Un monitor de máquinas virtuales (VMM) también conocido como **hypervisor**



# Xen es un monitor de MV que...

- Da soporte a múltiples MV con SO de uso común.

- Linux 2.4



- Windows XP
  - NetBSD



# Xen es un monitor de MV que...

- Permite instanciar MV de forma dinámica y utiliza un esquema de pago por utilización de recursos
- Se busca instanciar hasta 100 MV

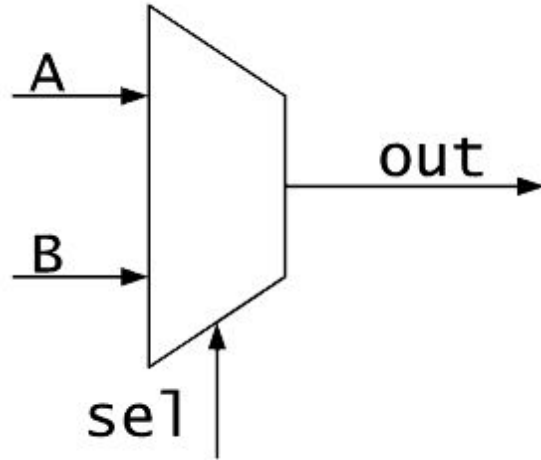
# Problemas de una implementación ingenua

- No se logra tener un comportamiento aislado de las MV
- Las MV interfieren entre sí en procesos como
  - Calendarización
  - Accesos a memoria
  - Tráfico de red
  - Accesos a disco



# ¿Cómo lograr aislamiento?

- Multiplexar el hardware
- Turnar a los SO
  - Más pesado que multiplexar procesos
  - Se logra aislar los SO

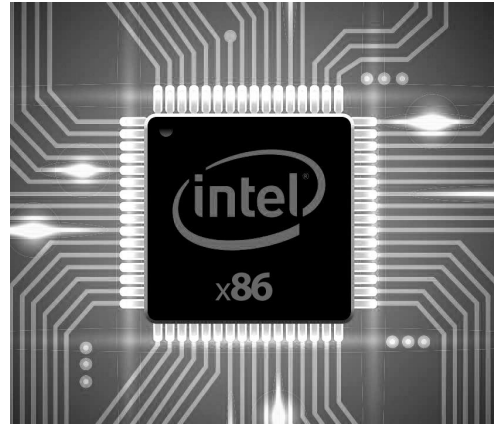


# La propuesta de Xen y sus características generales

1. Propuesta de Xen
2. Interfaz de la máquina virtual (manejo de memoria, de CPU y de E/S)
3. El costo de hacer portable un SO a Xen
4. Administración y Control

# Xen se enfoca en la arquitectura x86

- La arquitectura x86 no brinda soporte a virtualización total.
- Las tablas de paginación se manejan a nivel hardware.



Virtualizació  
n

Total

**VS**

Para-virtualización

# Virtualización Total

## Características

- Presenta una abstracción total del hardware que es idéntico al físico.
- Se pueden utilizar las versiones de los SO sin modificaciones.

## Contras (en x86)

- Si el VMM no tiene privilegios suficientes, las instrucciones de administración de recursos no se ejecutan.
- Costoso generar trampas en la MV para lograr que las instrucciones del VMM se ejecuten adecuadamente
- Difícil virtualizar la unidad de manejo de memoria en x86

# Paravirtualización

## Características

- Se genera una abstracción limitada, pero similar del hardware para cada MV.
- Se puede tener un mejor manejo de los recursos y los privilegios por el VMM.

## Contras

- Se tiene que modificar el código de los SO, manteniendo la misma interfaz binaria de aplicaciones (ABI).

# Requerimientos del diseño

- Virtualizar **todas** las características del hardware utilizadas por las ABIs.
- Dar soporte a los SO para que puedan correr múltiples aplicaciones
- Utilizar el enfoque de paravirtualización
- Los SO conocen algunos efectos de la virtualización.

Denali

**VS**

Xen



## Denali

- Miles de MV en red.
  - Aplicaciones desconocidas y poco utilizadas
- No virtualiza todo lo que requieren las ABIs existentes.
- No multiplexea : SO no protegido con una sola aplicación.

## Xen

- 100 MV
  - Aplicaciones y servicios muy utilizados.
- Virtualiza lo que requieren las ABIs existentes para lograr dar soporte a SO utilizados.
- Multiplexea: SO reales con múltiples procesos

## Denali

- El VMM realiza todas las operaciones de paginación.
- Se virtualizan las direcciones de memoria

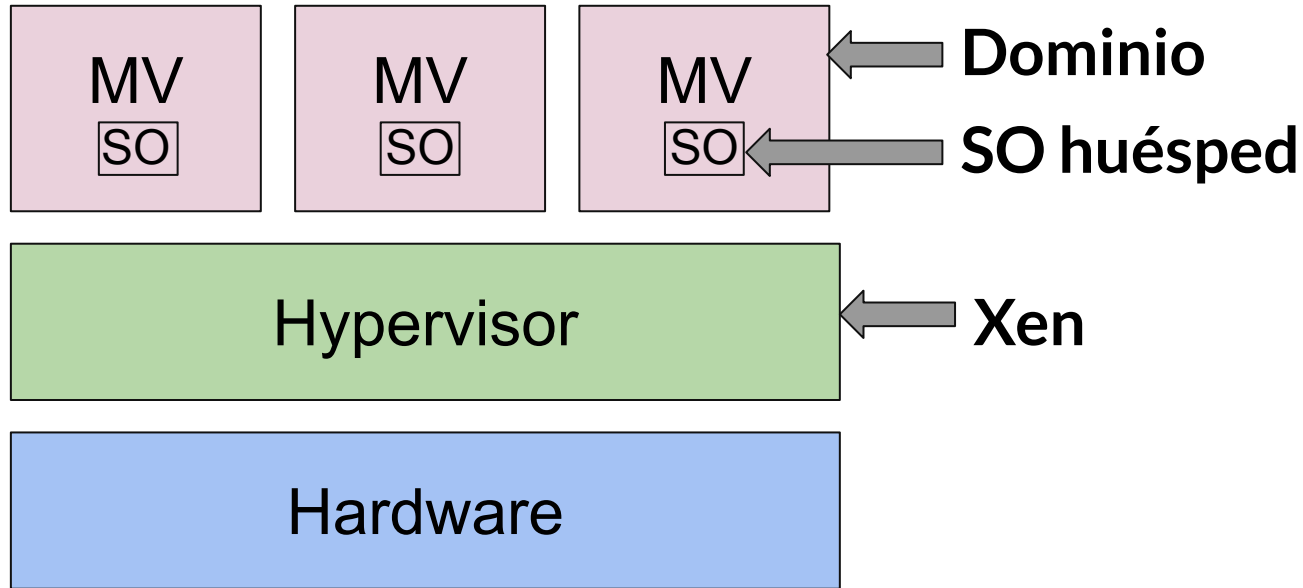


## Xen

- Xen asigna un espacio de memoria a las VM y estas se deben encargar de la paginación.
- Se muestran las direcciones de memoria y Xen sólo restringe el acceso.



# Hypervisor, Dominio y SO huésped

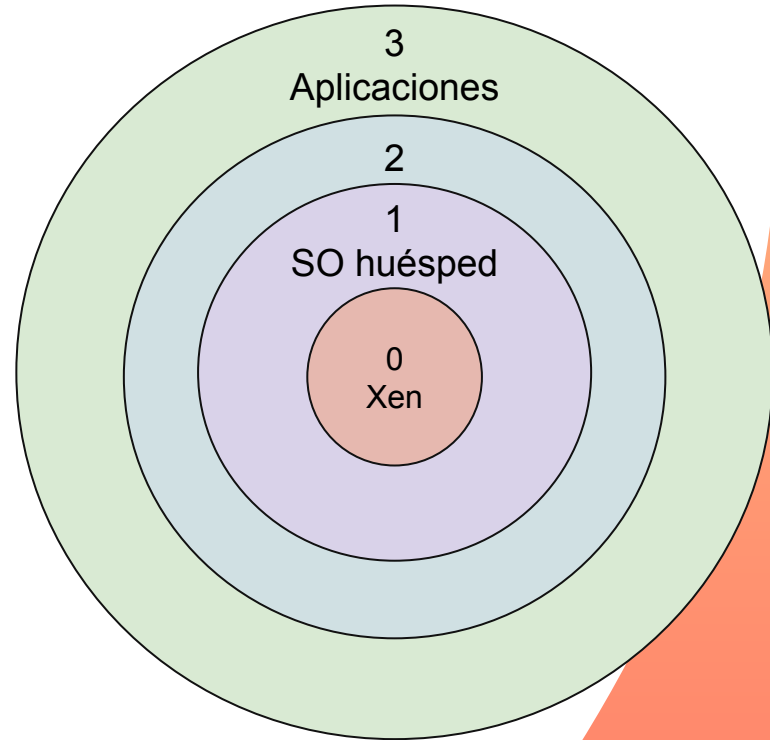
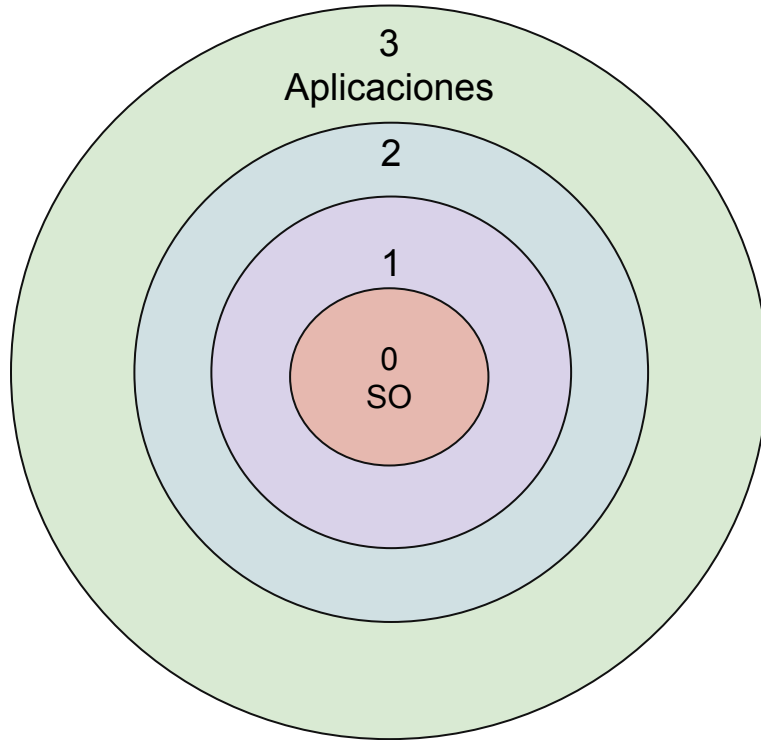


# Interfaz de la Máquina Virtual

# Manejo de Memoria

- Los SO huéspedes se encargan de asignar y administrar sus tablas de paginación.
- Xen verifica el comportamiento adecuado.
- Xen se encuentra en los 64 MB al inicio de cada dirección de memoria.

# CPU: Se modifican los niveles de privilegios



# CPU: Excepciones



- En su mayoría se manejan de forma idéntica que en el hardware.
  - La pila de excepciones no se modifica.
  - La excepción de fallo de página se muestra en otra pila en vez de leerse del registro CR2.
- Propagación de excepciones.

# Dispositivos E/S

- Los dispositivos se muestran de forma transparente.
- Los datos pasan a través de Xen haciendo uso de un espacio de memoria compartida administrada en un buffer anillado. (Más adelante lo veremos a detalle)





# **El costo de hacer portable un SO a Xen**

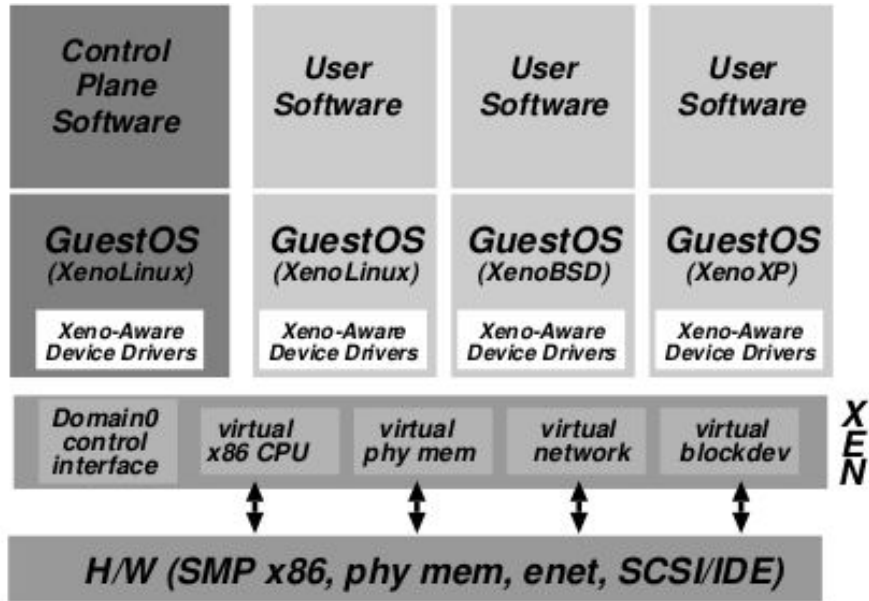
# Costo en líneas de código

OS subsection	# lines	
	Linux	XP
Architecture-independent	78	1299
Virtual network driver	484	–
Virtual block-device driver	1070	–
Xen-specific (non-driver)	1363	3321
<b>Total</b>	<b>2995</b>	<b>4620</b>
(Portion of total x86 code base	<b>1.36%</b>	<b>0.04%</b> )

**Table 2: The simplicity of porting commodity OSes to Xen. The cost metric is the number of lines of reasonably commented and formatted code which are modified or added compared with the original x86 code base (excluding device drivers).**

# Administración y Control

# Dominio 0

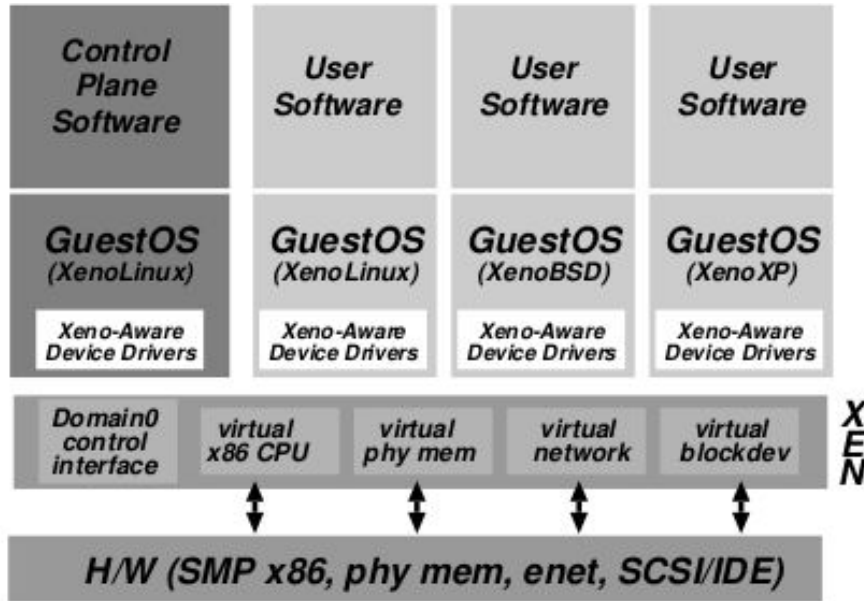


Xen no se encarga de la administración de MV a alto nivel....

Para eso está el  
**Dominio 0**

**Figure 1:** The structure of a machine running the Xen hypervisor, hosting a number of different guest operating systems, including *Domain0* running control software in a XenoLinux environment.

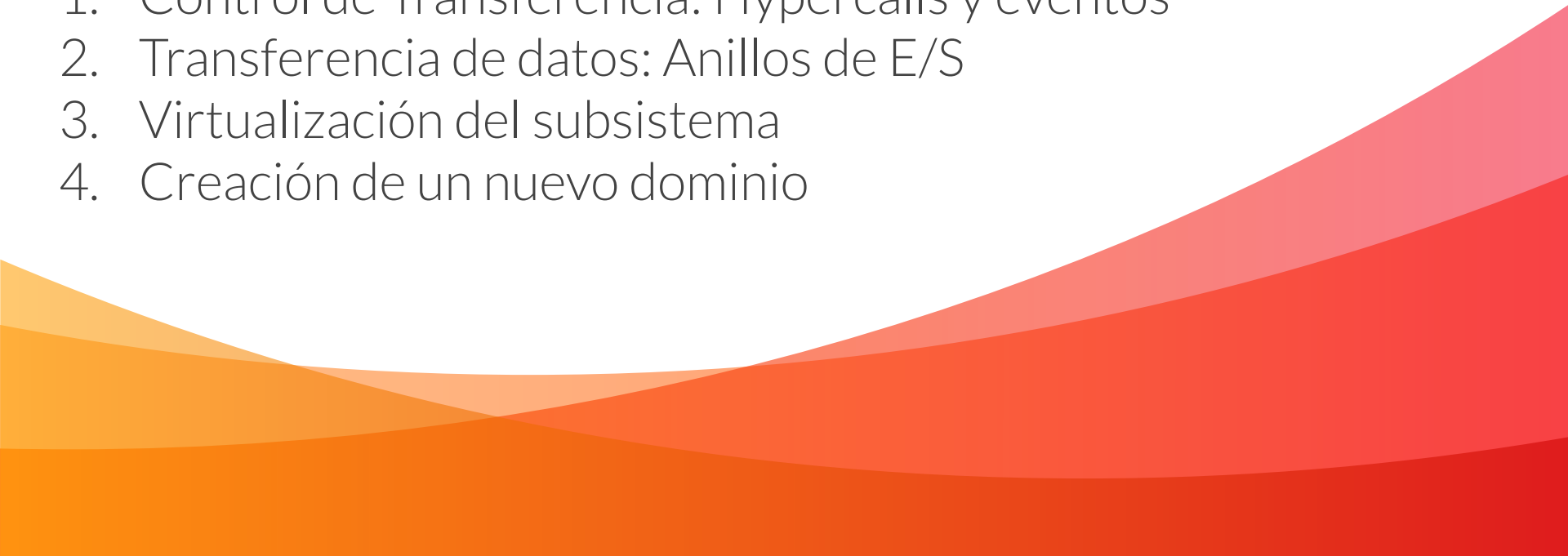
# Dominio 0



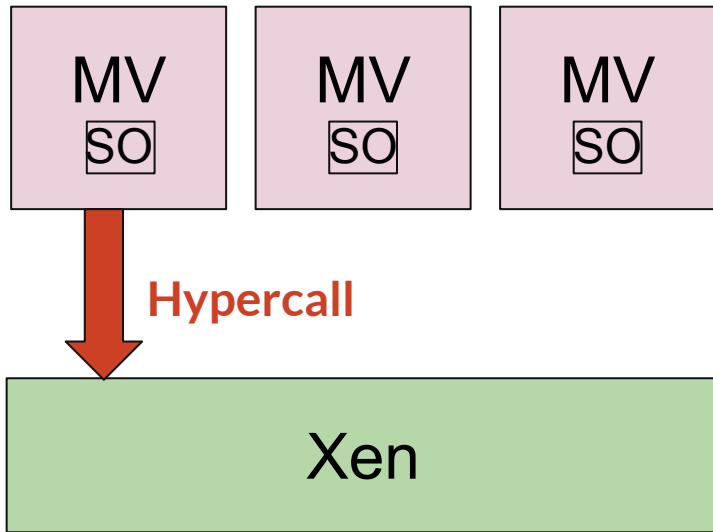
- Tiene acceso a la interfaz de control
- Crea y elimina otros dominios
- Controla los parámetros de calendarización de los dominios.
- Asigna recursos de memoria y red.
- Define bloques virtuales

**Figure 1:** The structure of a machine running the Xen hypervisor, hosting a number of different guest operating systems, including *Domain0* running control software in a XenoLinux environment.

# Diseño detallado de Xen

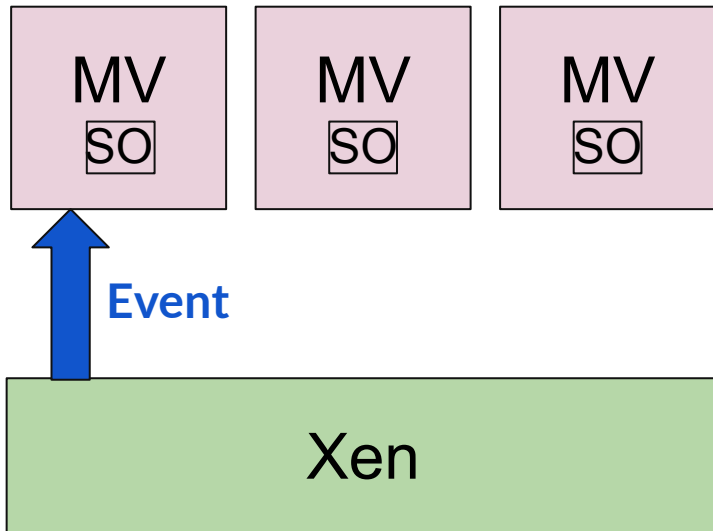
1. Control de Transferencia: Hypercalls y eventos
  2. Transferencia de datos: Anillos de E/S
  3. Virtualización del subsistema
  4. Creación de un nuevo dominio
- 

# Control de Transferencia: Hypercalls



- Síncronos
- Solicitan ejecutar operaciones privilegiadas en Xen
- Equivalentes a llamadas a sistema

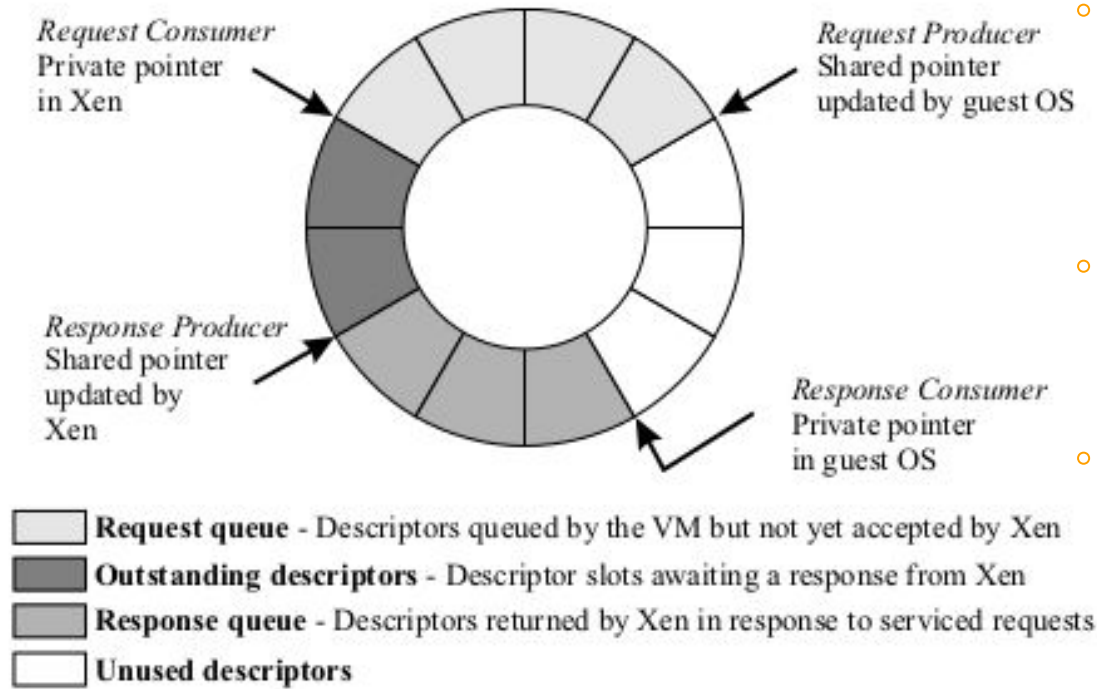
# Control de Transferencia: Eventos



- Asíncronos
- Se utilizan para dar notificaciones importantes.
- Funcionan de forma similar a las interrupciones.



# Transferencia de datos: Anillos de E/S



- Los descriptores hacen referencia a los buffers de E/S en cada SO huésped.
- Sólo Xen y el SO en cuestión tienen acceso al buffer de E/S.
- Xen puede reordenar las solicitudes.

**Figure 2: The structure of asynchronous I/O rings, which are used for data transfer between Xen and guest OSes.**

# Virtualización del Subsistema

# CPU: Calendarización de dominios

- Se utiliza el algoritmo Borrowed virtual time (BVT).
  - Se ejecuta el dominio que tenga menor tiempo virtual efectivo.
  - Un dominio puede tomar tiempo prestado del futuro para ejecutarse primero.

# Tiempo y Relojes

- Tiempo real
- Tiempo virtual
- Reloj de pared



# Traducción de direcciones virtuales

- Xen registra las tablas de paginación de un SO huésped con MMU.
- El SO huésped tiene libre acceso a la lectura a las tablas de paginación.
- Xen verifica y ejecuta las actualizaciones a las tablas de paginación.

# Traducción de direcciones virtuales

A cada marco de página se le asigna un valor:

- Directorio de página (PD)
- Tabla de paginación (PT)
- Tabla de descriptores local (LDT)
- Tabla de descriptores global (GDT)
- Editable (RW).

# Memoria Física

- Cuando se instancia una MV se le asigna la memoria física.
- Se utiliza la técnica de *ballooning*.
- La memoria asignada a una MV puede **no** ser contigua.





## Red

- Se presenta una abstracción de un firewall virtual.
- Al que se conectan las distintas interfaces virtuales de red (VIF) de las distintas MV.
  - Tienen 2 anillos para E/S
- Una lista de reglas que indica qué acciones tomar, la cual es administrada por el Dominio 0





## Red

- La transferencia de paquetes se encola en un buffer.
- Xen tiene acceso a los descriptores del buffer
- Se envían los paquetes usando round robin
- Para la recepción de paquetes el SO huésped intercambia páginas de memoria por los paquetes.

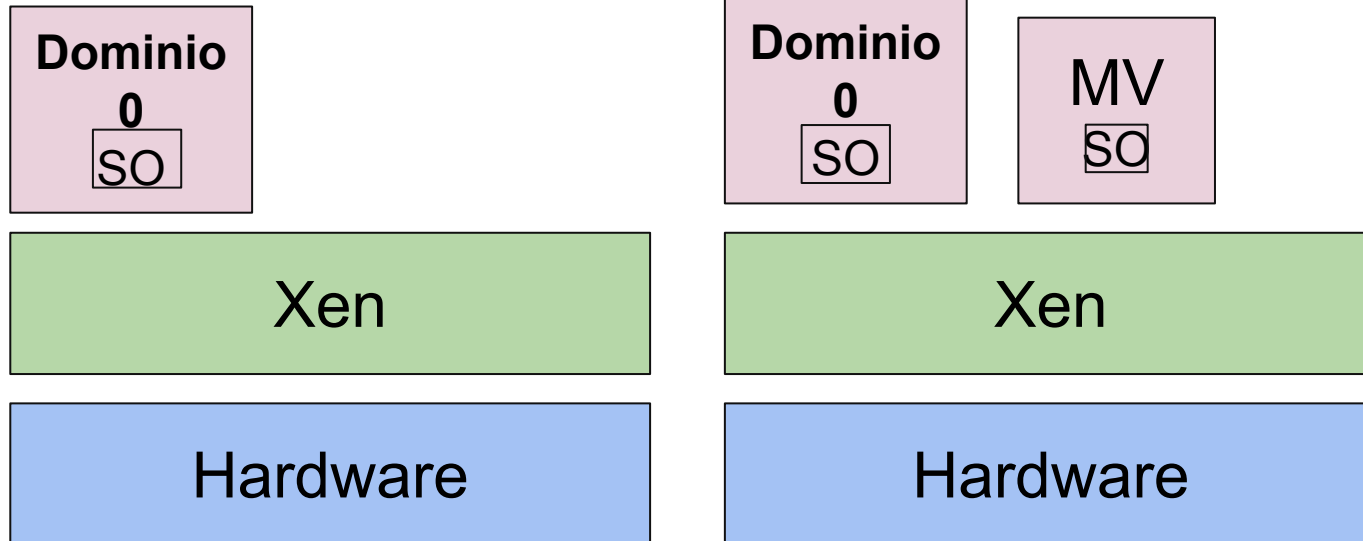


# Disco


- Sólo el Dominio 0 tiene acceso directo al disco.
- Para los demás bloques se presenta una abstracción por medio de dispositivos de bloques virtuales (VBD).
- El dominio 0 administra los VBD
- Xen participa en el reordenamiento de los accesos a disco

# Creación de un nuevo dominio

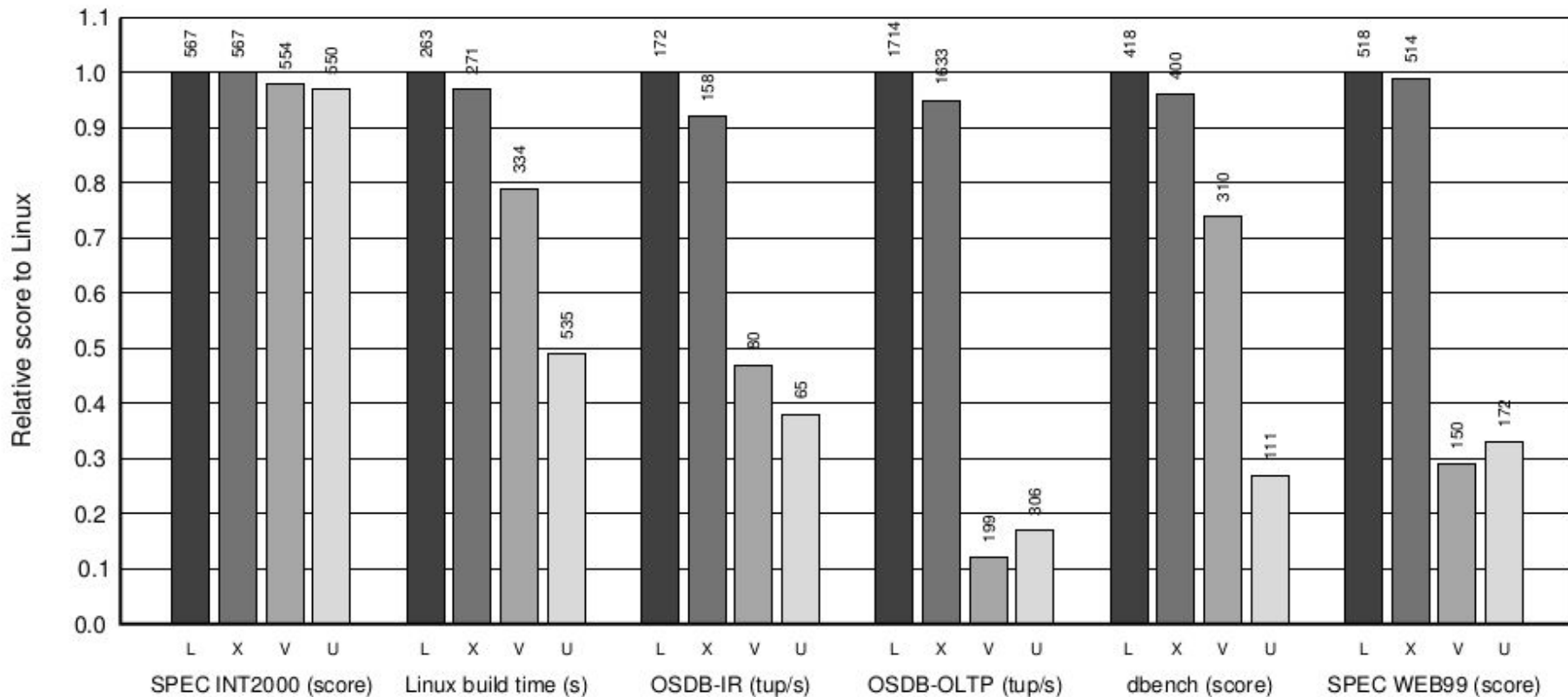
- El dominio 0 crea los nuevos dominios
- La dificultad en la creación de un dominio depende del SO huésped



# Evaluación

1. Desempeño Relativo
  2. Benchmarks en SO
  3. MV concurrentes
  4. Desempeño del aislamiento
  5. Escalabilidad
- 

# Desempeño relativo



# Benchmarks en SO

Config	null call	null I/O	stat	opens close	slct TCP	sig inst	sig hndl	fork proc	exec proc	sh proc
L-SMP	0.53	0.81	2.10	3.51	23.2	0.83	2.94	143	601	4k2
L-UP	0.45	0.50	1.28	1.92	5.70	0.68	2.49	110	530	4k0
Xen	0.46	0.50	1.22	1.88	5.69	0.69	1.75	<b>198</b>	<b>768</b>	<b>4k8</b>
VMW	0.73	0.83	1.88	2.99	11.1	1.02	4.63	874	2k3	10k
UML	24.7	25.1	36.1	62.8	39.9	26.0	46.0	21k	33k	58k

Table 3: **lmbench**: Processes - times in  $\mu s$

Config	2p 0K	2p 16K	2p 64K	8p 16K	8p 64K	16p 16K	16p 64K
L-SMP	1.69	1.88	2.03	2.36	26.8	4.79	38.4
L-UP	0.77	0.91	1.06	1.03	24.3	3.61	37.6
Xen	<b>1.97</b>	<b>2.22</b>	<b>2.67</b>	<b>3.07</b>	<b>28.7</b>	<b>7.08</b>	39.4
VMW	18.1	17.6	21.3	22.4	51.6	41.7	72.2
UML	15.5	14.6	14.4	16.3	36.8	23.6	52.0

Table 4: **lmbench**: Context switching times in  $\mu s$

Config	0K File		10K File		Mmap	Prot	Page
	create	delete	create	delete	lat	fault	fault
L-SMP	44.9	24.2	123	45.2	99.0	1.33	1.88
L-UP	32.1	6.08	66.0	12.5	68.0	1.06	1.42
Xen	32.5	5.86	68.2	13.6	<b>139</b>	1.40	<b>2.73</b>
VMW	35.3	9.3	85.6	21.4	620	7.53	12.4
UML	130	65.7	250	113	1k4	21.8	26.3

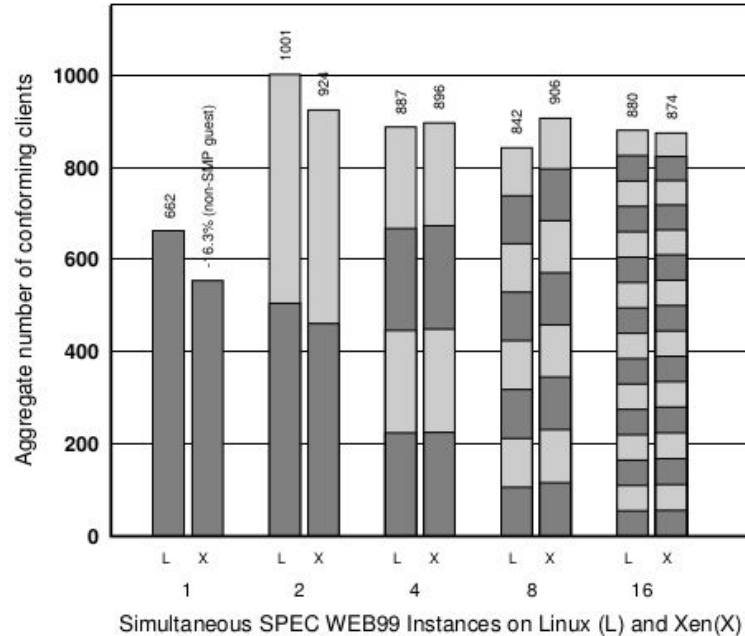
Table 5: **lmbench**: File & VM system latencies in  $\mu s$

# Desempeño en red

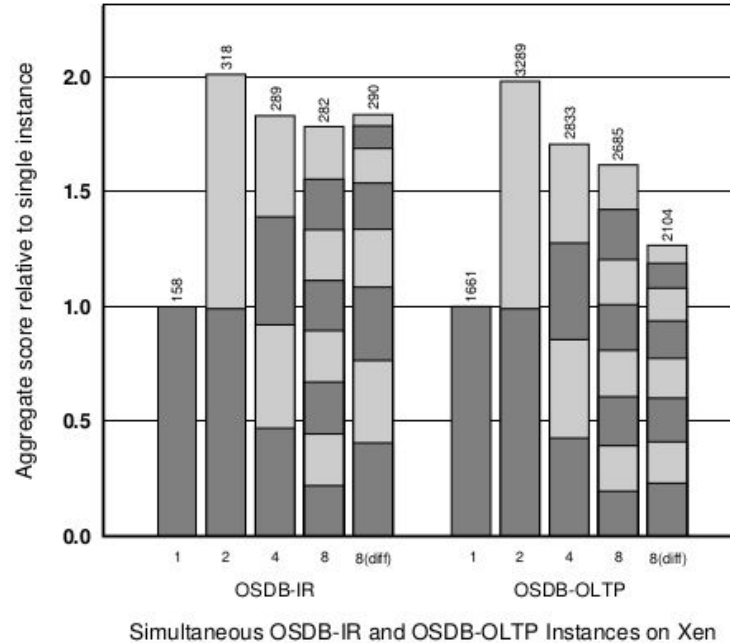
	TCP MTU 1500		TCP MTU 500	
	TX	RX	TX	RX
Linux	897	897	602	544
Xen	897 (-0%)	897 (-0%)	516 (-14%)	467 (-14%)
VMW	291 (-68%)	615 (-31%)	101 (-83%)	137 (-75%)
UML	165 (-82%)	203 (-77%)	61.1(-90%)	91.4(-83%)

**Table 6: `ttcp`: Bandwidth in Mb/s**

# MV concurrentes



**Figure 4: SPEC WEB99 for 1, 2, 4, 8 and 16 concurrent Apache servers: higher values are better.**



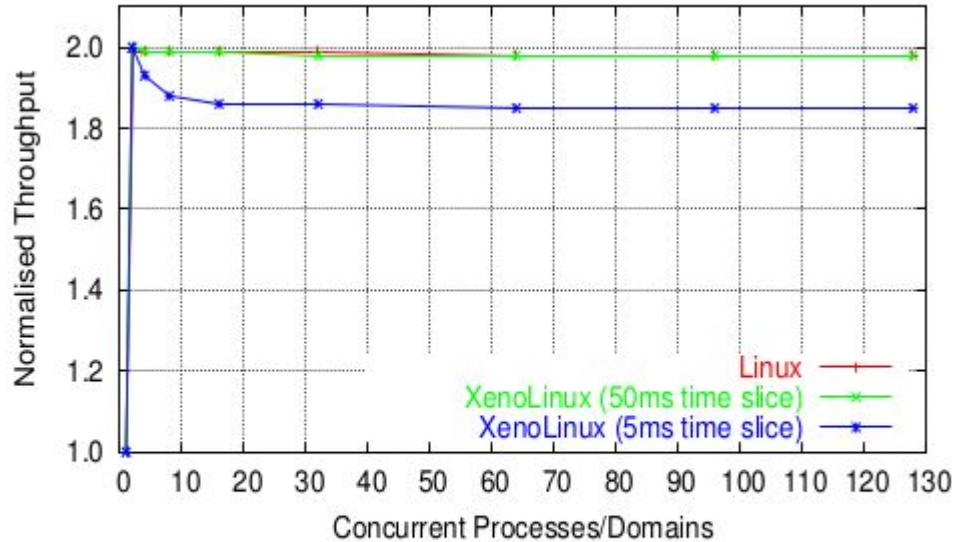
**Figure 5: Performance of multiple instances of PostgreSQL running OSDB in separate Xen domains. 8(diff) bars show performance variation with different scheduler weights.**



# Desempeño del aislamiento

- Correr 4 dominios
  - a. PostgreSQL/OSDB-IR
  - b. SPEC WEB99
  - c. Creación intensiva de pequeños archivos en directorios grandes y *disk bandwidth hog*
  - d. *Fork bomb* y asignar 3GB de memoria virtual, esperar un error, liberar todas las páginas y repetir
- Sólo hubo una pérdida de 4% y 2% en el desempeño

# Escalabilidad



**Figure 6: Normalized aggregate performance of a subset of SPEC CINT2000 running concurrently on 1-128 domains**

# Trabajo Futuro

The background features abstract, flowing shapes in shades of orange and red. On the left, there are overlapping orange waves. On the right, there are overlapping red waves that transition into a lighter pink area at the top right corner.

# Trabajo Futuro

- Aumentar la eficiencia de los bloques con un cache universal de buffer.
- Last-chance page cache
- Mayor control de E/S para poder establecer un precio adecuado
- Generar reglas que eviten comportamiento antisocial en la red
- Trabajar sobre XenoxP

# Actualmente

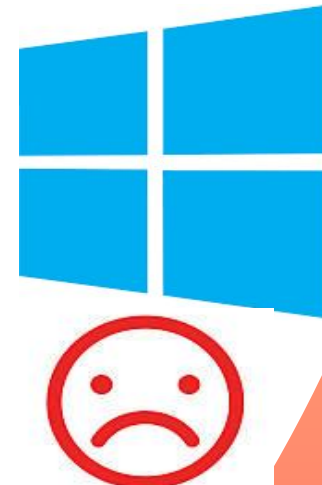
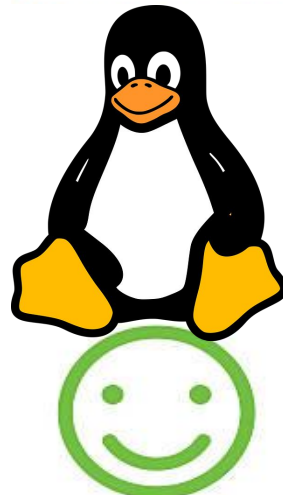
- Paravirtualización
- Virtualización desde el hardware
- Arquitecturas x86 y ARM

Distribution	run as PV guest
Alpine Linux 2.3.x	yes
Alpine Linux 2.4.x	yes
Arch Linux (x86-64 only [5])	yes
CentOS 5	yes
CentOS 6	yes
Debian 5.0 (Lenny)	yes
Debian 6.0 (Squeeze)	yes
Debian 7 (Wheezy)	yes
Fedora 14	yes
Fedora 16	yes
OpenEmbedded 1.3+	yes
OpenSUSE 11.4	yes
Oracle Linux 5	yes[2]
Oracle Linux 6	yes[2]
RHEL 5	yes [3]
RHEL 6	yes
SLES 10	yes
SLES 11	yes
Ubuntu 10.04	yes[1]
Ubuntu 11.04	yes[1]
Ubuntu 12.04	yes[1]

# Actualmente

Distribution	run as PV guest	PVHVM
FreeBSD >=10	No[1]	Yes
NetBSD >=3	Yes[2]	No
OpenBSD 5.9	no[6]	yes[3]
OpenIndiana[4]	no[4]	no
Solaris 11	yes[5]	no

Distribution	run as PV guest
Windows Vista	no
Windows XP	no
Windows 2008	no
Windows 2003	no
Windows 2000	no
Windows 7	no



[https://wiki.xen.org/wiki/DomU\\_Support\\_for\\_Xen](https://wiki.xen.org/wiki/DomU_Support_for_Xen)

# Críticas

The background of the slide features abstract, flowing shapes in shades of orange and red. On the left side, there are overlapping wavy shapes in various tones of orange. On the right side, there are similar wavy shapes in shades of red and pink, creating a sense of movement and depth.

# Críticas

- Asegurar que al hacer ballooning se limpien las páginas
- Buscar dar soporte a multiprocesadores
- Asegurar el buen funcionamiento del Dominio 0.
- Privacidad de los datos de las MV
- ¿En los experimentos con un solo dominio consideran un dominio adicional al 0?