El Imperio Convergente

DC + ISP bajo un mismo control

Ariel S. Weher
ariel@ayuda.la

ayuda.LA

Ayuda.LA

Qué hacemos

- Ingeniería de redes L2/L3
- Plataformas de virtualización y cloud privadas
  - Proxmox, KVM, Ceph, Contenedores
- Observabilidad y automatización
  - NetOps, monitoreo, pipelines de logs/telemetría
- Seguridad y autenticación
  - AAA, RADIUS/TACACS, Zero-Trust
- Troubleshooting avanzado e incident response

Cómo trabajamos

- Foco en continuidad operativa y zero-downtime
- Diseño + ejecución + transferencia de conocimiento
- Integración con equipos técnicos del cliente
- Soluciones modernas, vendor-neutral y reproducibles

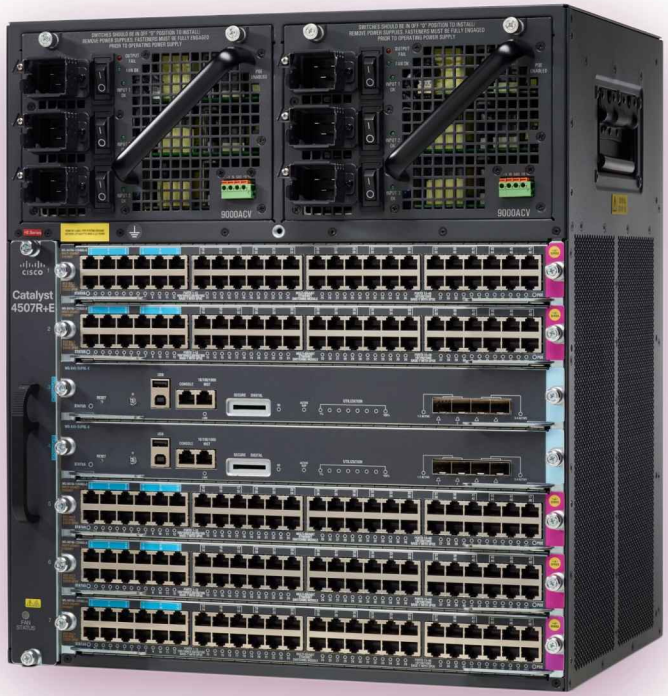
Nuestra Misión: resolver ingeniería real para redes que mueven tráfico real.

Erase una vez...
en una galaxia muy lejana...

AS27976 - Cooperativa de Morteros

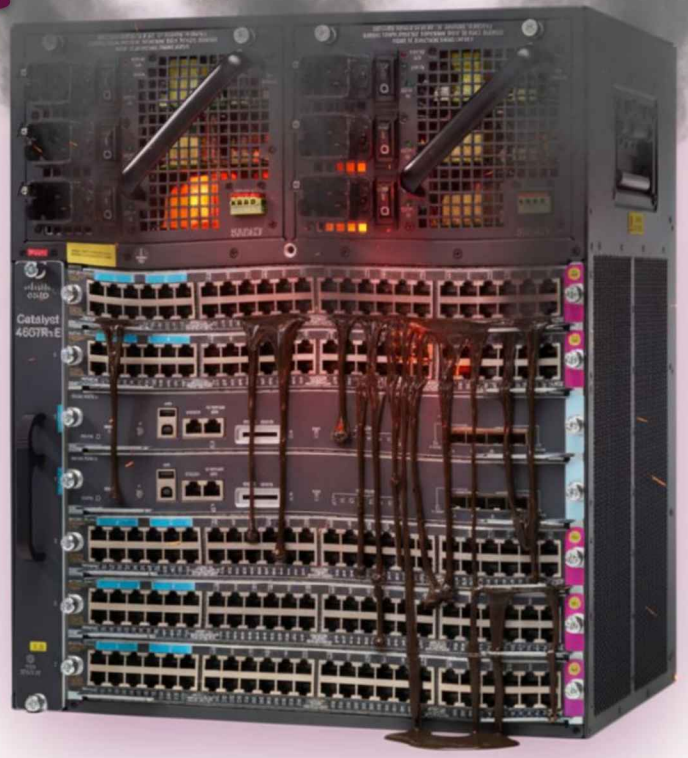
- Distribuidor eléctrico
- ~ 7.000 suscriptores de internet con DHCP
- IPTV Multicast + OTT
- IoT - OT
- Cloud Datacenter propio
- Desarrollo de aplicaciones y hardware

- ❑ Degradación de servicio inexplicable
- ❑ Cientos de VLANs heredadas
  - ❑ STP dependientes
- ❑ Switches Cisco legacy
  - ❑ 2800W en 11U y punto único de fallo
- ❑ Software de virtualización licenciado
- ❑ Monitoreo y visibilidad de eventos limitada



La herencia de la operación diaria

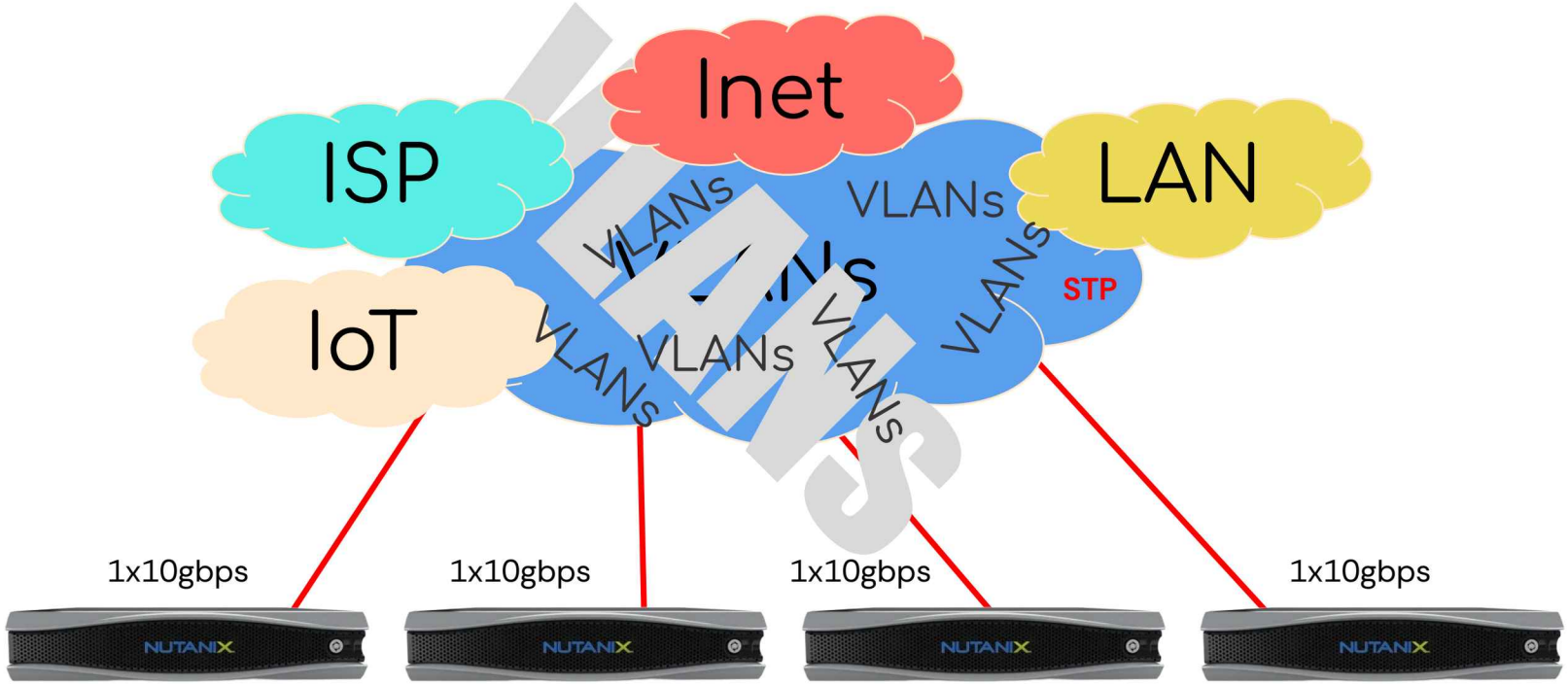
- No existía una fuente de verdad
- Plan de direcciones un un excel
- Documentación parcial
- Accounting limitado de las acciones de los suscriptores
- 250+ VLANs configuradas manualmente
- Spanning Tree defaulteado
- Multicast defaulteado
- Multi-vendor sin estrategia: Cisco, MikroTik
- Recursos de hardware insuficientes para la operación diaria (TCAM / FIB)
- Sistema de monitoreo parcial con alertas solo ante fallos graves
- Plan de acción estríctamente reactivo ante la llamada de los clientes
- Múltiples areas de la empresa manejando la red



La herencia del hypervisor

- Vendor lock-in fuerte y ecosistema cerrado
  - Costo de licencias elevado
  - Obliga a comprar hardware certificado y mantenerlo actualizado
- Migración compleja desde y hacia otras plataformas
- Curva de aprendizaje alta que requiere skills específicos
- Escalabilidad costosa en donde CPU/RAM/Storage crece en bloque y con hardware específico
- Menos flexibilidad para arquitecturas personalizadas
- Limitaciones con OSS y menor libertad técnica
- Riesgo estratégico por contexto del vendor/mercado

El problema real del hypervisor



¿Qué hicimos?

La red tiene que funcionar siempre

Esto no se negocia.

Se debe poder sobrevivir al menos al fallo de un equipo de core.

Los datos siempre deben estar disponibles y seguros

No solo resistir la caída de un disco, sino de un nodo de almacenamiento completo.

La información de la empresa debe sobrevivir a las personas

Nada se debe hacer sin estar planificado y documentado previamente.

IPv6 es transversal a todos los proyectos

Cada proyecto debe prever su funcionamiento en una red que incluso pueda ser basada solo en IPv6.

Plan de Acción

- Se limitaron accesos a los equipos de red
- Se implementó un monitoreo completo de todos los puntos donde se pudieran tomar mediciones
- Se auditó cada dispositivo y cada configuración
- Se estableció una sola fuente de verdad
- Se prohibió la creación de nuevos servicios hasta dejar la red estable y encontrar la causa de los cortes
- Se iniciaron los procedimientos de compra de acuerdo a las políticas internas del cliente, con nuestro equipo formando parte en las negociaciones con los proveedores seleccionados

Análisis de eventos pre y post

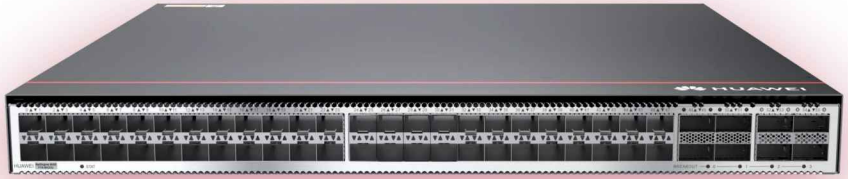
- Se estableció una política clara de entender lo que pasa buscando de apuntar a los procesos y no a las personas.
- Se utiliza un modelo de IA entrenado para monitorear logs y predecir fallos.
- Cada vez que hay un evento se hace un informe post mortem.

Informe Post Mortem

1. Breve introducción del incidente o error
2. Línea temporal
3. Análisis de la causa
4. Impacto y daños causados
5. Acciones tomadas para solucionar el incidente
6. Medidas para que no ocurra otra vez
7. Lecciones aprendidas

Equipos adquiridos

- 4 x Huawei NE8000 F1A
  - c/licencia Advanced p/ L3VPN
- 4 x Huawei S6730 línea H
- 2 x Cisco Nexus 9300
- 2 x Fortigate 400F
- Servidores
  - De acuerdo al ciclo de compra habitual



Eficiencia

- 40X capacidad de forwarding
- 1/20 consumo eléctrico

Bases del proyecto

- Layer 3 everywhere
  - *undo portswitch* en todos lados.
  - No más dependencia de los distintos sabores de STP.
- Decidimos aprovechar los features incluidos en las plataformas pagando la menor cantidad de licencias posibles.
  - Esto nos llevó a elegir soluciones con MPLS por sobre VXLAN, EVPN o SRv6.
- Se comenzó a gestar una red en paralelo pensada desde cero con multipath.
  - Escalabilidad horizontal garantizada con ECMP.

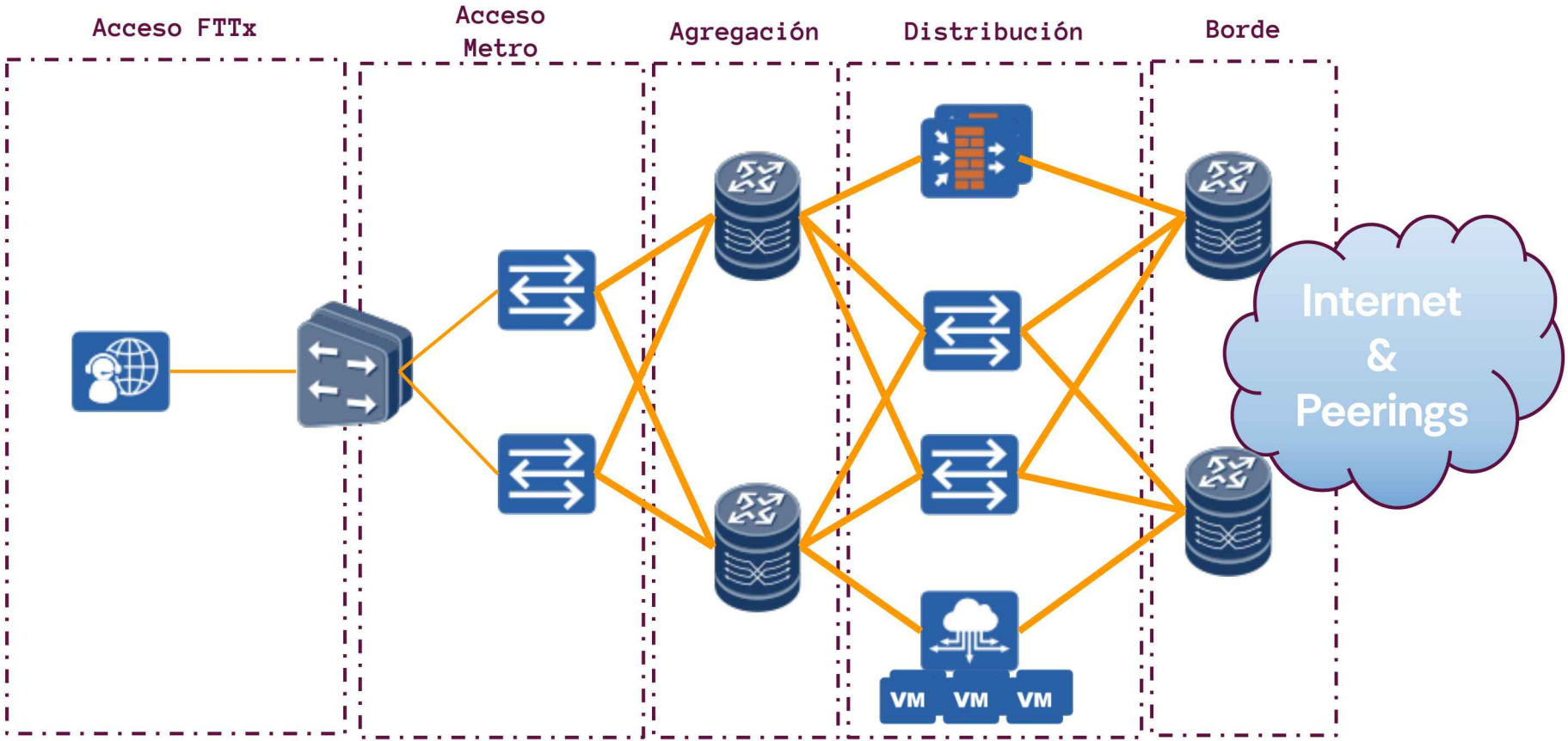
- Open source first, no más vendor lock-in
- Maximizar los esquemas de protección de datos de la Cooperativa y sus clientes.
- Autenticación centralizada en todos los equipos (TACACS+ / RADIUS)
- Migración desde DHCP a PPPoE Dual Stack para todos los suscriptores FTTx

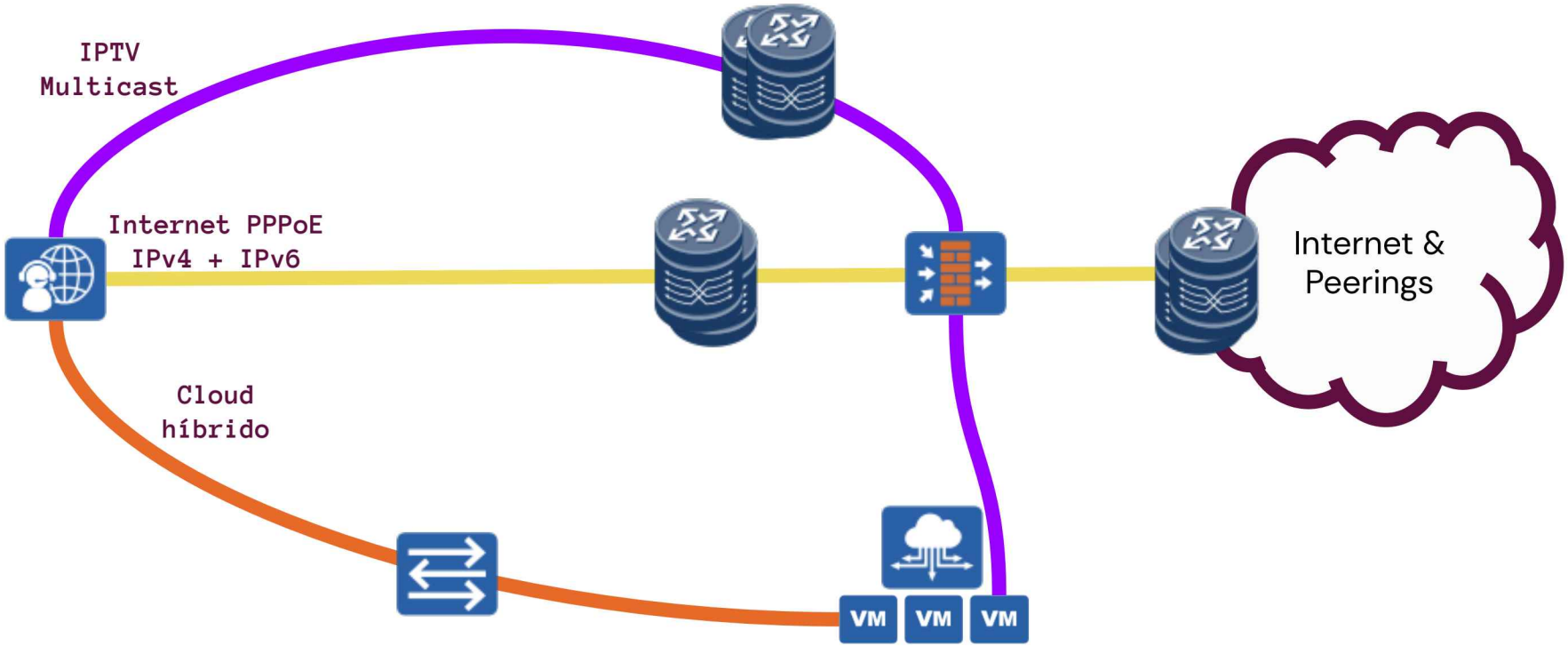
MPLS ya no es una utopía

- **Para el underlay:** IS-IS en todos los equipos.
- **Para el overlay:** BGP y EVPN donde se soporte sin costo extra.
- **BFD** en todos los equipos compatibles.
- Label distribution: LDP automático.
- MPLS VPNs: Segmentación de servicios.
- Ingeniería de tráfico donde corresponda.
- Soporte out-of-the-box de RPKI, FlowSpec, etc.

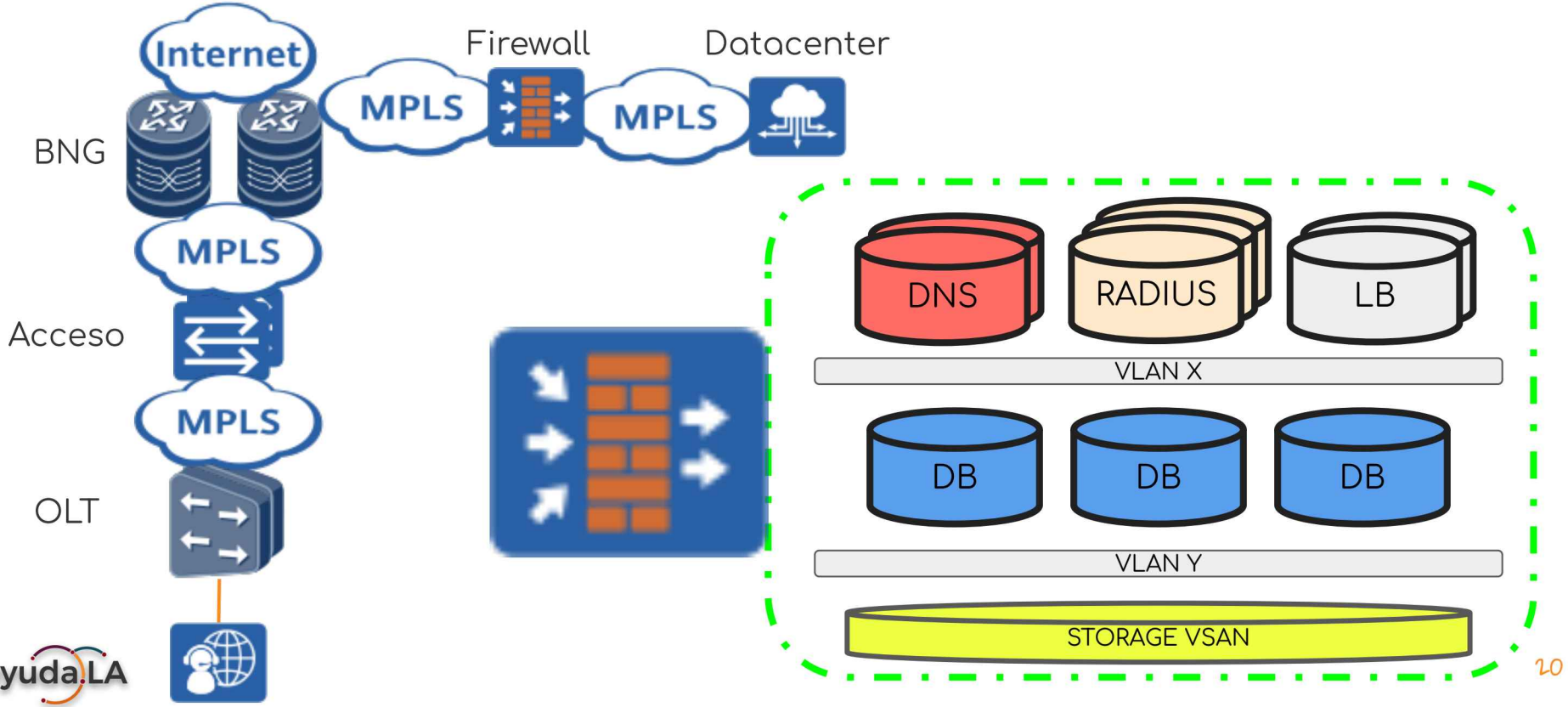
Resultados cuantificados:

- Outages: CERO
- Optimización en la utilización enlaces redundantes.
- Mejora de tiempos de convergencia ante una falla:
  - 30-50 segundos → < 1 segundo.


Acceso FTTx
Acceso Metro
Agregación
Distribución
Borde
Internet & Peerings
VM
VM
VM


IPTV
Multicast
Internet PPPoE
IPv4 + IPv6
Cloud
híbrido
VM
VM
VM
Internet &
Peerings

Despliegue de Servicio de BNG



NetBox: La fuente de la Verdad

- Repositorio único y confiable con la información real de la infraestructura.
- Define cómo debe ser la red, no solo cómo está hoy.
- Evita configuraciones manuales, planillas sueltas e "info en la cabeza del técnico".
- Base para automatizar, validar y documentar.
- “DECCE” Documenta → Etiqueta → Configura → Conecta → Enciende.

¿Por qué hay que implementar NetBox?

- Diseñado para redes complejas (ISP / DC / empresas).
- Inventario centralizado: dispositivos, IPs, BGP, VRFs, VLANs, servicios.
- API completa → se integra con Ansible / Nornir / GitOps.
- Evita errores humanos y configuraciones inconsistentes.
- Permite generar configs automatizadas basadas en datos reales.
- Open Source, moderno y orientado a automatización.
- **Resultado:** menos errores, más velocidad, red coherente y automatizada

La receta para la provisión de servicio a los clientes

Capacidades de la red

- Hasta 32K sesiones redundadas
- Line rate forwarding
- Soporte de COA
- Dual Stack nativo
- Accounting diferenciado
- Transporte end to end con MTU 9K

Capacidades del DC

- Redundancia física
- Orquestación de VMs
- Redundancia de storage
- Redundancia de acceso a red
- Monitoreo contínuo predictivo y proactivo

Servicios redundados

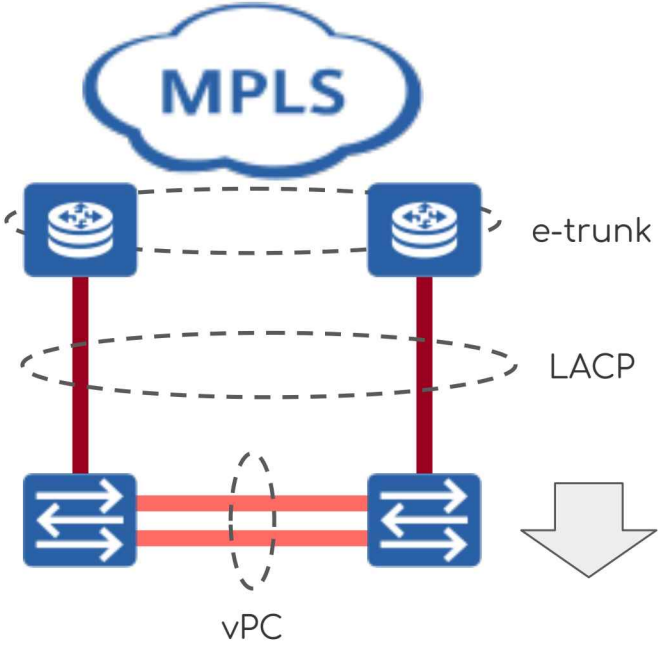
- Servicios replicados
- Balanceadores de datos
- Firewall L7
- Monitoreo de latencias
- Sistemas creados in-house según estrictos criterios internos de calidad

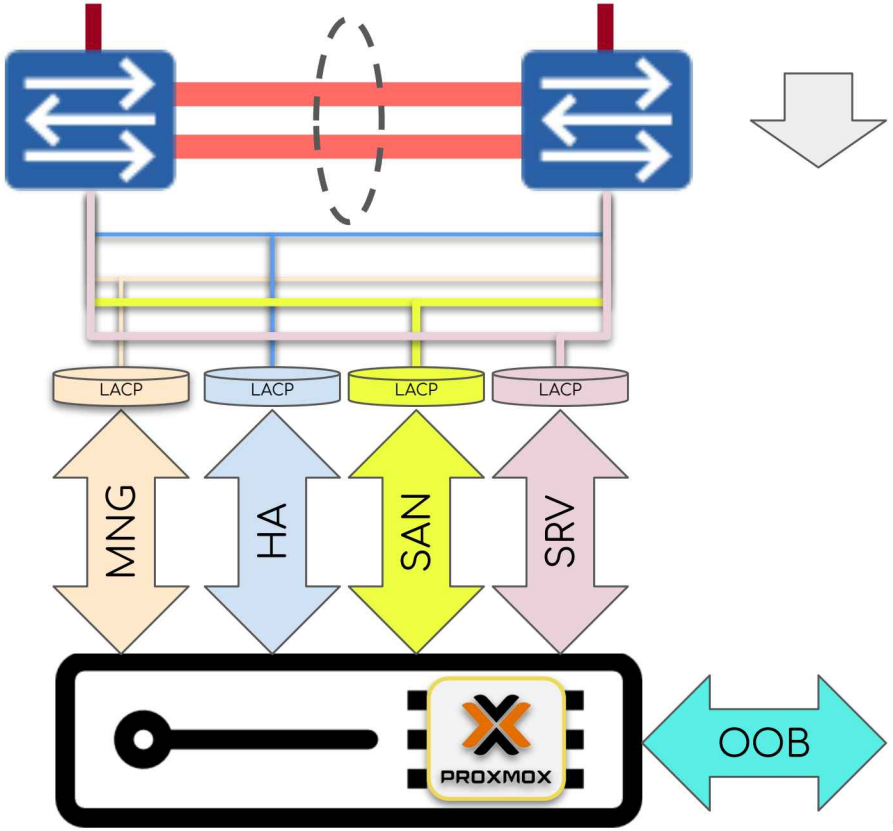
Proxmox VE 9

- Proxmox VE v9 consolida un stack carrier-grade open source.
- Mejora la disponibilidad, automatización y escala.
- Con costos accesibles que pueden ser cero en caso de optar por el soporte de comunidad.
- Almacenamiento distribuido nativo
  - CephFS (archivos)
  - RBD (bloques)
  - Object Gateway (S3-like)
- Rendimiento alto y paralelo
- Tolerancia a fallos
  - (fallos de nodos/discos sin perder datos)

- **Fabric SDN integrado**
  - Soporte para topologías Leaf-Spine
  - Descubrimiento y gestión más simple de mallas L2/L3
  - Pensado para entornos distribuidos y alta disponibilidad
- **Multipath automático**
  - Balanceo y failover transparente entre caminos
- **VLAN / QinQ avanzado**
  - Segmentación flexible a nivel tenant / VNF
  - Uso típico: VLAN WAN, VLAN clientes, transporte L2 mayorista
- **Gestión centralizada de redes virtuales**
  - Bridges, bonds, VXLAN, VLAN, rutas
- **Compatibilidad con NIC de 25/40/100G**
  - DPDK-ready

Interconectando correctamente los servers





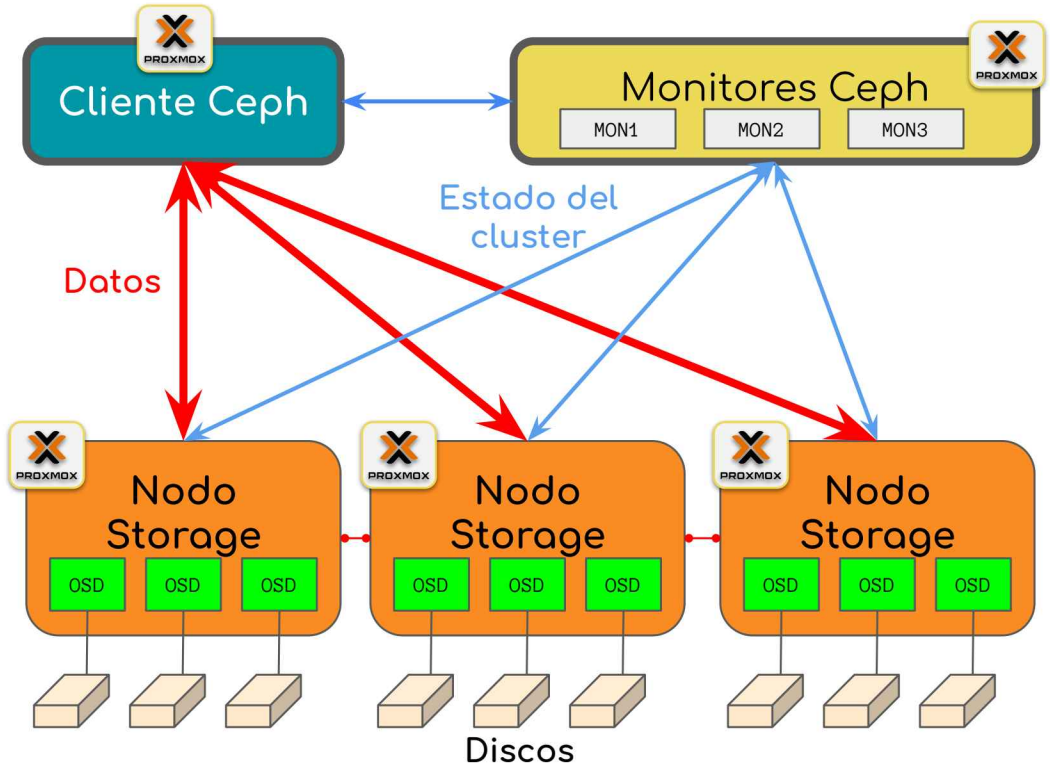
Configuración de OpenFabric

```
hostname pvenodo1
!
interface loopback_of
 ip address 172.20.255.41/32
 ipv6 address fd32:beba:cafe::1/128
 ip router openfabric pve1
 ipv6 router openfabric pve1
 openfabric passive
exit
!
interface ens1f4
 ip router openfabric pve1
 ipv6 router openfabric pve1
exit
!
interface ens1f4d1
 ip router openfabric pve1
 ipv6 router openfabric pve1
exit
!
router openfabric pve1
 net 49.0001.1720.2025.5041.00
exit
```

```
hostname pvenodo2
!
interface loopback_of
 ip address 172.20.255.42/32
 ipv6 address fd32:beba:cafe::2/128
 ip router openfabric pve2
 ipv6 router openfabric pve2
 openfabric passive
exit
!
interface ens1f4
 ip router openfabric pve2
 ipv6 router openfabric pve2
exit
!
interface ens1f4d1
 ip router openfabric pve2
 ipv6 router openfabric pve2
exit
!
router openfabric pve2
 net 49.0001.1720.2025.5042.00
exit
```

“VSAN” en Proxmox VE 9

- Arquitectura distribuida y sin punto único de falla
  - Datos replicados o erasure-coded entre nodos
  - Continúa funcionando incluso con fallas de discos o servidores
- Escala horizontal real (scale-out)
  - Sumás nodos/OSDs y crece la capacidad y el rendimiento
  - Ideal para crecimiento orgánico en cooperativas e ISPs regionales
- Rendimiento alto y consistente
  - Lecturas/escrituras paralelas a través del cluster
  - Soporte para NVMe, SSD, HDD, jerarquización y cache
- Flexibilidad total
  - Hardware abierto (no vendor lock-in)
  - Multi-datacenter, topologías híbridas y edge computing
- Alta disponibilidad sencilla
  - No requiere cabinas SAN ni redes FC costosas
  - Perfecto para clusters de producción y POPs distribuidos



OpenFabric

```
pvenodo1# show openfabric neighbor
Area openfbrc:
 System Id           Interface   L  State        Holdtime SNPA
 pvenodo2            ens1f4      2  Up           10       2020.2020.2020
 pvenodo2            ens1f4d1    2  Up           10       2020.2020.2020

pvenodo2# show openfabric neighbor
Area openfbrc:
 System Id           Interface   L  State        Holdtime SNPA
 pvenodo1            ens1f4      2  Up           10       2020.2020.2020
 pvenodo1            ens1f4d1    2  Up           9        2020.2020.2020

root@pvenodo2:~# ip route
[...]
172.20.255.41 nhid 54 proto openfabric src 172.20.255.42 metric 20
     nexthop via 172.20.255.41 dev ens1f4 weight 1 onlink
     nexthop via 172.20.255.41 dev ens1f4d1 weight 1 onlink
```

Cosas que aprendimos

(algunas a los golpes)

Pequeños problemas se vuelven bolas de nieve

- Ante un evento, generalmente se aconseja ampliar la capacidad de los equipos.
- Estas ampliaciones solo oculta el problema hasta que cobra el tamaño suficiente como para necesitar “otro upgrade”.
- Las inversiones deben planificarse y ejecutarse con información y fundamentos.
- Siempre buscar en la información de logs y monitoreo para ver si la necesidad del upgrade no se evita al solucionar una mala configuración o un servicio que falla.

El MTU de la red es crítico

- 9216 fue en nuestro caso el default adecuado
  - Cuidado con los tramos alquilados a terceros
  - MTU IPv4 != MTU IPv6
- Cada encapsulamiento tiene sus requisitos
  - El header MPLS quita 4 bytes y se pueden perder hasta 12 + los adicionales
  - El header VXLAN quita 8 bytes y se pueden perder hasta 50 + los adicionales
- Distintas marcas lo miden diferente, cuidado.

Siempre comprar el soporte oficial

- Los equipos fallan.
- Los sistemas operativos se deben actualizar.
- Corregir vulnerabilidades es necesario.
- Los manuales no explican todo.
- Las MAC address a veces son todas iguales.
- Los clientes a veces necesitan features que no están documentados.
- “La guerra no se puede poner en pausa para ir a comprar balas”

Cuidado con los switches...

Aunque nos parezcan "caros", no todos los switches son adecuados para cualquier posición en la red.

Modelo	Tipo de buffer	Tamaño de buffer	Notas técnicas
Huawei CloudEngine 6870-48S6CQ-EI	Deep/Shared	4 GB	Buffer compartido profundo, ideal para microbursts y tráfico exigente
Huawei CloudEngine S6730H-48	Shared/Deep	16 MB	Buffer compartido, optimizado para ráfagas y tráfico convergente
Cisco Nexus 9300	Shared (Dynamic)	50 MB (standard) / 1 GB (deep buffer SKU)	50 MB modelos normales, 1 GB modelos "deep buffer"; buffer compartido/dinámico
Arista 7050SX (ej: 7050SX2-128)	Shared (Multi-level)	12 MB (estándar) / 16 MB (SX2)	Buffer principal compartido, dividido entre pools dedicados y compartidos

En nuestra experiencia reciente, el mejor equipo que se desempeñó como "switch" siempre fue el *router* Huawei NE8000 F1A

Virtual Output Queue

VoQ optimiza el procesamiento interno de los switches permitiendo que el tráfico fluya hacia múltiples destinos, aun si alguno está temporalmente congestionado, mejorando la eficiencia y el control sobre el tráfico en redes avanzadas de alto rendimiento.

Fabricante	Familia / Modelo	Notas clave
Huawei	CloudEngine XH16800, CloudEngine 12800, CloudEngine 16800	Arquitectura cloud/disaggregated soportando VoQ en switching fabric
Arista	7280X/7280R/7280R3, 7500R, 7800R, 7050X3, 7050X4	Diversos modelos fixed/modular con VoQ, ideales para topologías CLOS
Juniper	QFX10000, PTX Series, algunos modelos EX (core/distribución)	VoQ en Packet Forwarding Engine (PFE) y disaggregated fabric
Broadcom-like	Muchos modelos OCP y "white box" (Edgecore, Dell EMC S6100, etc.)	Depende del chip Broadcom incorporado (Trident, Tomahawk, etc.)

Para finalizar...

El imperio convergente está al alcance

Arquitectura ISP + DC: Principios Clave

- **Arquitectura Correcta:** No tiene que ver con equipos costosos.
- **Open Source:** Ofrece las mismas funcionalidades sin la carga de licencias.
- **Visibilidad:** Es la prioridad para la mejora continua (No se puede mejorar lo que no se mide).
- **Plan de Migración:** Implementación por fases para garantizar cero interrupciones.
- **Inteligencia Artificial:** Permite una operación más dinámica y profunda a un precio más conveniente.
- **Sostenibilidad:** Modelo basado en expertise remoto complementado con visitas regulares.

¡Jubilemos esos switches de core!

Q&~A

ayuda.LA