DISEÑO Y DESPLIEGUE DEL CENTRO DE PROCESO DE DATOS DE RESPALDO DEL MINISTERIO DE LA PRESIDENCIA

Julián Hernández Vigliano
Jefe de Área de Comunicaciones
Subdirección General de Sistemas de Información
Ministerio de la Presidencia

Coautor:
Juan Carlos Gómez Garzón
Jefe de Servicio de Sistemas de Información
Subdirección General de Sistemas de Información
Ministerio de la Presidencia

Palabras clave

CPD de respaldo
Virtualización
Extensión de vlans
Clustering
HSRP

Resumen de su Comunicación

En este documento se recogen las propuestas de diseño y despliegue del centro de respaldo en el Ministerio de la Presidencia. Se enfoca desde dos grandes puntos de vista: los sistemas (datos y servicios) y las comunicaciones que permiten que los anteriores sean accedidos en caso de fallo del sistemas principal o de manera paralela en función de los servicios.

1 Introducción

El Centro de Proceso de Datos de respaldo del Ministerio de la Presidencia se despliega con la idea de incrementar la tolerancia a fallos del sistema informático a niveles muy altos. Consiste en duplicar los servicios y los datos fundamentales para el funcionamiento del Ministerio en un CPD paralelo situado en un edificio distinto al del CPD principal.

El Ministerio de la Presidencia cuenta con un CPD desde el cual presta los servicios de aplicaciones, almacenamiento, comunicaciones y de proceso de datos al personal de Ministerio y sus organismos autónomos. A pesar de que este centro cuenta con las procedimientos de seguridad puestos en marcha para asegurar su funcionamiento y los sistemas y electrónica cuentan con diversos grados de duplicidad para garantizar el servicio, se ha detectado la necesidad de contar con un centro alternativo desde el cual prestar los mismos servicios de manera simultánea en caso de que problemas generalizados tuviesen lugar en cualquiera de los dos centros.

Se escoge como ubicación física del nuevo CPD un edificio, de los varios en los que se distribuye el Ministerio, que ya tenía un espacio lo suficientemente grande y acondicionada para alojar servidores con sistema eléctrico y acondicionamiento de aire duplicado, UPS, etc. En el se alojan tres racks de servidores, dos racks de comunicaciones y un rack con el sistema de almacenamiento.

2. Objetivos

Los objetivos del proyecto son los siguientes:

DESDE EL PUNTO DE VISTA DE SISTEMAS

1. Incrementar la tolerancia a fallos del sistema de almacenamiento. La finalidad del proyecto es que una restauración masiva de backup sea una posibilidad remota.

2. Incrementar la tolerancia a fallos de todos los servicios informáticos fundamentales. Se pretende que en caso de que uno de los CPD quede inutilizable por alguna razón (corte de suministro eléctrico, obras, catástrofe…) el otro sea capaz de proporcionar todos los servicios informáticos al Ministerio sin que los usuarios lo noten.

3. Disponer de un sistema de rápido de recuperación ante desastres de los sistemas del Ministerio de la Presidencia. El objetivo es también agilizar la recuperación ante un desastre, en caso de que la duplicación de los datos en el Centro de Respaldo no sirva para mantener el servicio (en caso de corrupción de datos).

4. Disponer de un sistema de backup más ágil y de mayor capacidad. Mayor rapidez de recuperación e incremento de la capacidad de respaldo de la organización.

5. Facilitar la duplicación y copia de seguridad de sistemas informáticos, no sólo de datos. El objetivo es que la caída de un servidor físico no suponga la interrupción del servicio que presta ese servidor, que otros servidores sean capaces de mantener el servicio.

6. Consolidación de sistemas informáticos mediante la virtualización con VMWare Enterprise. Reducción considerable de máquinas físicas, con el consiguiente ahorro energético que ello supone.

DESDE EL PUNTO DE VISTA DE COMUNICACIONES

1. Independencia de la localización física del cpd de respaldo. La topología y el diseño del centro de respaldo debe ser invariable ante las posibles ubicaciones alternativas que se puedan barajar para la colocación inicial / cambio futuro.
2. Robustez ante múltiples fallas en ambos centros
3. No complicar la gestión del equipamiento
4. Mantener la infraestructura de direccionamiento.Arquitectura de Nivel 3 solo ligeramente modificada.
5. La arquitectura debe permitir implantarse por fases, coexistiendo el cpd principal con el progresivo desarrollo del nuevo cpd de respaldo.
6. Flexibilidad de operación (Activo/Activo o Activo/Pasivo): Soportar esquemas de funcionamiento desde el punto de vista de respaldo de sistemas (datos y servicios) tanto activo/pasivo como activo/activo, pudiendo atender simultáneamente las peticiones de los usuarios (Activo/Activo) o bajo la variante que solo uno de los equipos tenga el control total del tráfico mientras que el redundante se espera alguna caída para entrar en operación (Activo/Pasivo)
7. Todas las redes externas al ministerio deben ser alcanzables por todos los usuarios tanto externos como internos en ambos centros.

3. Características principales y detalles de funcionamiento de respaldo de datos y servicios

3.1 Respaldo de datos

La replicación de datos utilizada tiene una configuración Activo-Pasivo. La unidad de almacenamiento de respaldo está inactiva, sólo recibiendo la replicación de datos, y a la espera de que un Administrador de Sistemas bascule la replicación y la haga cambiar de rol y convertirse en activa.

En principio se pensó dejar la decisión de cual era el CPD activo en un momento de crisis en manos de un servidor que continuamente estuviese preguntando por el estatus de las unidades de almacenamiento masivo de ambos edificios. En caso de detectar un fallo en la unidad de almacenamiento del CPD activo, el sistema automáticamente tomaría la decisión de bascular los recursos al CPD de respaldo. Esta opción se descartó por dos razones:

- Implicaba alojar un tercer CPD de otro edificio y tirar dos líneas de fibra desde ese nuevo edificio a cada uno de los Centros de Proceso de Datos.
- Podría darse el caso de que los dos CPD estuviese a pleno rendimiento, sin ningún tipo de problema y que, por alguna razón, la fibra que uniría el servidor de control con el CPD principal se cortase. En ese caso el sistema tomaría la decisión errónea de bascular los recursos. Esto supondría perdida de conectividad de los usuarios, y posiblemente la pérdida de algún documento, sin ninguna necesidad, pues en realidad el sistema no estaba caído.

Se optó por la opción de dejar la decisión en manos de un Administrador de Sistemas. Las unidades de almacenamiento masivo utilizadas permiten la realización de clones periódicos de toda la información. La puesta en marcha de un clon como sustituto del disco de producción en caso de necesidad sería mucho más rápida que una restauración desde una copia de seguridad.

- Hardware implicado en la replicación

  Se utiliza el sistema de almacenamiento masivo tipo SAN HP StorageWorks EVA 4000. Hay un sistema de este tipo en cada CPD unidos por varias líneas de fibra óptica.

  Cada CPD dispone de cuatro cabinas con 6 TB brutos en discos de fibra y 6 TB brutos en discos FATA. El software para gestionar la EVA 4000 es el HP Command View.

- Software implicado en la replicación
  - Business Copy de HP. Este software sirve para realizar los clones. Se hace un clon todas las noches de los discos fibra a los discos FATA.

  - Continuous Access. Software para gestionar la replicación de datos de una EVA 4000 a la otra. Se define una replicación síncrona entre ambas EVAs (los datos se escriben simultáneamente en las dos EVAs). Todos los datos están continuamente replicados en ambos CPDs, por si se diera un desastre en uno de ellos. Están definidos en total 12 grupos de replicación (un grupo diferente por cada instancia de SQL y Exchange, uno para cada disco de quórum de los cluster físicos, uno para los archivos de usuarios, otro para los archivos de aplicaciones y otro para las máquinas virtuales).

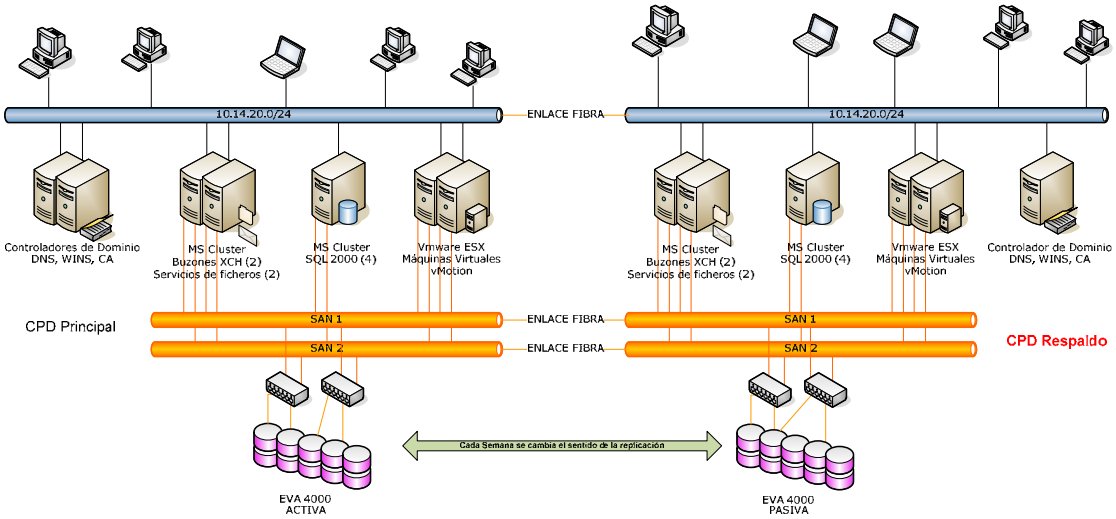
  - Replication Solution Manager. Software para realización de scripts para gestionar la EVA 4000. Todas las semanas, mediante un script programado a la 4 de la mañana de los lunes, se hace un balanceo de los recursos de la SAN de un CPD a la del otro. De forma que las dos EVAs cambian de rol y pasan de activa a pasiva y viceversa.

De esta manera garantizamos tanto que funciona el balanceo entre ellas, como que ambas están totalmente operativas.

El script de balanceo consta de tres partes:

1. PreFailOver
   - Suspende todas las máquinas virtuales
   - Pone Offline todos los Grupos de Cluster de Windows
   - Para el Servicio de Cluster en todos los nodos de Cluster Windows
2. FailOver
   - Cambia EVA 4000 Activa, hace el failover de discos
3. PostFailOver
   - Resetea todos los nodos de Cluster Windows con lo que se ponen On Line todos los Grupos de Cluster Windows. De esta forma los nodos de cluster reconocen sin problemas los nuevos discos presentados.
   - Se hace un reescan de discos en los servidores VMWARE
   - Se reinician las máquinas virtuales suspendidas

ARQUITECTURA FÍSICA DE ALMACENAMIENTO



3.2 Posibles incidencias en la red SAN de almacenamiento masivo:

FALLOS en una EVA 4000.

- Fallo de 1 o 2 discos FC simultáneamente:
  - Protección Raid de discos ◊ se sustituyen los discos y se regenera la información automáticamente.
  - Ni los sistemas ni los datos se ven afectados.
- Fallan 3 o más discos FC a la vez
  - Automáticamente se desactiva la EVA para que no se corrompan los datos de la Eva de Respaldo.
  - Fallan los sistemas, pero no afecta a los datos
  - Hay que balancear manualmente la EVA.
- Fallo en una controladora
  - La otra controladora asume todo el trabajo.
  - Se sustituye en caliente
  - No afecta ni a los sistemas ni a los datos.
- Fallan las dos controladoras
  - Fallan los sistemas pero los datos no son afectados
  - Hay que balancear manualmente la EVA.

FALLO en el RACK completo de la EVA 4000 (apagón, incendio …)

- No existe corrupción de datos
  - Fallan los sistemas y no afecta a los datos
  - Balanceo manual de EVA activa

- Se produce una corrupción de datos
  - Necesaria restauración de Backup:
    - Clon Disco FATA EVA. (Todas las noches)
    - Copia de Respaldo a Disco SCSI. (Copia diaria, Semanal)

- Copia de Respaldo a Cinta. (Diaria, Semanal, Histórico)

3.3 Respaldo de servicios

- Infraestructura física.

  - 20 servidores HP BL460 XEON DUAL con 2 procesadores y 10 Gb de memoria (10 en cada CPD) con VMWare ESX alojando máquinas virtuales
  - 17 servidores HP proliant de diferentes modelos repartidos entre ambos CPDs con algún tipo de redundancia (NLB o Cluster Service) o reparto de roles en caso de controladores de dominio
  - Dos HP Proliant (uno en cada CPD) con todo el software necesario para la replicación y manejo de las controladoras de disco y los scripts de balanceo de recursos de un CPD a otro y de realización de clones.

- Software para garantizar la tolerancia a fallos:

  Servicio de Cluster Físico de Microsoft:
  - 1 cluster de 4 nodos de Correo y Servicio de Ficheros (2 nodos en cada CPD)
  - 1 cluster de 2 nodos de SQL (1 nodo en cada CPD)

  Cluster NLB (Network Load Balanced) de Microsoft
  - 4 NLB físicos con un nodo en cada Centro de Respaldo
  - 5 NLB virtuales con un nodo en cada Centro de Respaldo

  Máquinas virtualizadas con VMWARE ESX con posibilidad de balancear de un CPD a otro mediante Virtual Center.
  Tres Domain Controllers (2 en el CPD principal y 1 en el CPD de respaldo).

  - Servicios respaldados:
  - Correo.- Tanto correo de Internet como correo interno, acceso http al correo y correo para PDAs
  - Bases de datos
  - Servidores de archivos, tanto los archivos de usuarios como los archivos de aplicaciones
  - Servidores Proxy de acceso a Internet.
  - Controladores de dominio
  - Aplicaciones Web
  - MOSS 2007
  - Servicio Web de mensajería a teléfonos móviles
  - Servicio de acceso remoto a aplicaciones (CITRIX).
  - Live Communication Server
  - Servicio de Windows Media Server
  - Servicio de actualizaciones automáticas de Microsoft (WSUS)
  - Servicio de resolución de nombres DNS y WINS

3.4 Nuevo Sistema De Backup

- Sistema de Copias de respaldo basado en el producto software DATA PROTECTOR de Hewlett-Packard.

- Dicho sistema se implementa en servidores HBA con conexión a la red SAN y robot de cintas que triplica la capacidad del anterior.
- Ventajas Operativas:
  - Concurrencia de proveedor HW/SW
  - Aprovechamiento y optimizacion de trafico en red de almacenamiento SAN.
  - Gestion de la copia de respaldo por el servidor propietario de la misma.
  - Posibilidad de SnapClon en "caliente".

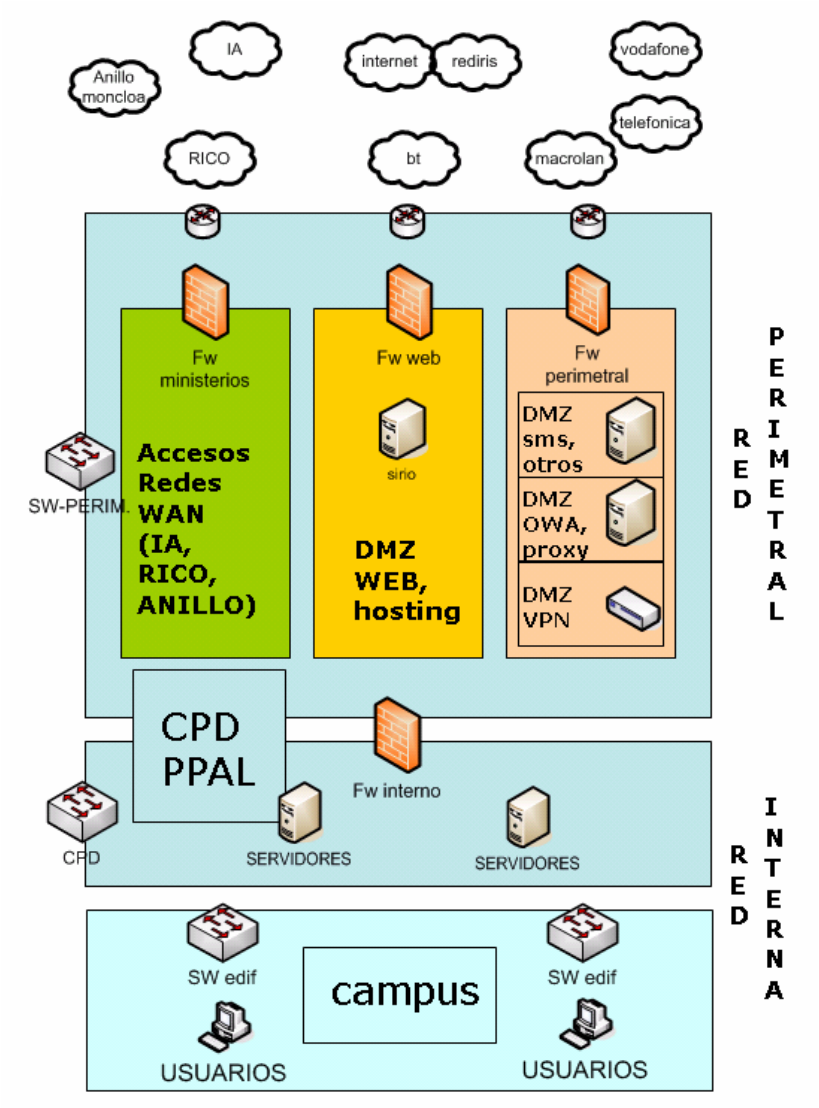
4. ASPECTOS DE DISEÑO DE COMUNICACIONES

Aproximadamente la mitad de usuarios del Ministerio están repartidos en el resto de Ministerios, de ahí que la importancia de las comunicaciones externas en el cpd de respaldo sea tan imporante, puesto que si sólo se duplica la capacidad operativa para los usuarios internos del complejo, sólo estaremos solucionando la mitad del problema.
Todas las comunicaciones hacia / desde el exterior están concentradas en la electrónica correspondiente en el CPD principal. Las comunicaciones internas están enfocadas a unir los conmutadores de cada edificio con el conmutador central del CPD principal, el cual recoge las redes de usuarios y servidores.
Cualquier fallo en el CPD principal puede afectar tanto a las comunicaciones internas (los usuarios hacia los servidores) como externas (usuarios externos hacia los servidores del CPD y usuarios internos hacia redes exteriores).
Es por ello que, a la par de contar con un centro de respaldo que garantice la duplicidad y salvaguarda de los datos y la capacidad operativa del Ministerio, es necesario duplicar las comunicaciones hacia / desde el exterior.

En lo que respecta a las comunicaciones, éstas se pueden subdividir en:

- Comunicaciones desde el exterior:
  - Accesos de usuarios remotos por VPN, Citrix o correo webmail
  - Accesos a aplicaciones publicadas hacia Internet
  - Accesos desde los Ministerios y Organismos Autónomos por la red RICO para las tareas semanales de Consejo de Ministros
  - Accesos a aplicaciones desde la Intranet Administrativa
- Comunicaciones hacia el exterior
  - Navegación Internet
  - Correo electrónico
  - Intranet Administrativa
  - Red RICO
  - BOE
  - Anillo del Complejo de la Moncloa
- Comunicaciones internas
  - Aquellas que unen los diferentes edificios del Ministerio de la Presidencia (con sus usuarios y servidores (RRHH, Difusión multimedia, etc.) hacia el CPD principal en INIA.

4.1 Situación de Partida



Los puntos más importantes de la red son

• El Switch CPD se encarga de centralizar y agregar todo el tráfico de los usuarios del MPR, además de dar conectividad a los servidores del CPD

• El Firewall interno, es la primera barrera que tienen los usuarios para acceder a los recursos de la red, protegiendo a su vez a los usuarios de posibles accesos no controlados.

• El Firewall Ministerios contiene varias DMZ que protegen y filtran los accesos a las redes externas como los múltiples ministerios a los cuales el MPR tiene conectividad, y a la red del Anillo.

• El Firewall perimetral se encarga de gestionar los accesos y aplicar las políticas de seguridad con los accesos de redes externas a los servicios públicos del MPR, (VPN, Correo, Citrix, etc). Entre sus DMZ podemos encontrar el terminador de túneles y a varios servidores

• El Terminador de Túneles IPsec, es la pieza fundamental para el servicio de VPNs del MPR, además de ser a través de éste que las sedes de Galicia, Zaragoza, CIS y Consejo de Estado tienen acceso a la red del MPR.

• El Switch perimetral, se encarga centralizar y direccionar los paquetes provenientes de las diferentes zonas de la red del MPR que se encuentran detrás de los Firewalls antes comentados.

Para la solución de centro de respaldo se colocará para cada uno de estos elementos críticos de la red, su homólogo en dicho centro, en algunos casos es obligatorio que sea un elemento exactamente igual, y en otros casos existe la flexibilidad de utilizar un elemento distinto con tal que soporte las técnicas necesarias para ejercer la redundancia.

4.2. Criterios del Diseño

Solución Nivel 2

Todos los switches que actualmente están funcionando en la red del MPR, deben tener conectividad con el Centro de Respaldo para que los usuarios de la red al momento de una caída en el CPD principal tengan disponibilidad con todos los servicios. La idea consiste en tener dos topologías tipo estrella, una estrella tendría su centro en el switch del CPD principal y la otra estrella (la redundante) en el switch que se instalaría en el Centro de Respaldo. Entre ambos switches debe haber también conectividad directa y con suficiente ancho de banda para transportar el tráfico agregado de la red, por tal motivo se sugieren directamente enlaces 10GE.

Este switch de backup serviría incluso para darle acceso a los servidores en el centro de respaldo, de esta forma se seguirían los mismos lineamientos de topología que se tienen actualmente implementados en la red del MPR

Al momento de implementar esta redundancia aparecen en la red bucles, que si no son controlados traerían problemas al rendimiento de la red, por lo que es imprescindible la utilización de un protocolo de Spanning Tree. El STP comprueba continuamente el estado de la red mediante la transmisión de paquetes BPDU, de forma que la detección de un cambio en la topología (debida a un fallo en un enlace, inserción de nuevo equipamiento, etc) provoca una reconfiguración automática en los puertos, para prevenir la pérdida de conectividad y asegurar la eliminación de bucles. Dentro de las diversas modalidades de este tipo de protocolos destacamos dos:

- C 802.1d ó STP tradicional: este es el protocolo más básico de STP, y que incorporan la mayoría de los conmutadores. El tiempo de convergencia (tiempo que tarda la red en estar estable y operativa tras un cambio en la topología) de este protocolo es aproximadamente 50 segundos.
- C 802.1w ó RSTP (Rapid Spanning Tree Protocol): esta es una variante más actual del protocolo STP, y permite tiempos de convergencia inferiores a los dos segundos (normalmente no suele superar el segundo). Es completamente interoperable con 802.1d, aunque en los enlaces con electrónica corriendo el antiguo protocolo se pierde la rapidez de RSTP.

Evidentemente la segunda versión de STP, 802.1w, es mejor y más eficiente que el STP

Antiguo que es la que se recomienda en el proyecto. Para la construcción de la topología libre de bucles STP elige primero un nodo root, que va a ser la semilla a partir de la cual se definirá el árbol de STP. Por ello, siempre que se activa el protocolo STP, hay que configurar quiénes van a ser los conmutadores root primario y secundario. Estos nodos, por sentido común, serán los equipos centrales de la red, donde se va a concentrar la mayoría del tráfico. Actualmente el root de la red es el Switch del CPD, y cuando exista el Centro de Respaldo éste switch seguirá siendo el root principal, mientras que el switch de backup sería la segunda opción para ejercer ese rol en la red.

Solución de nivel 3

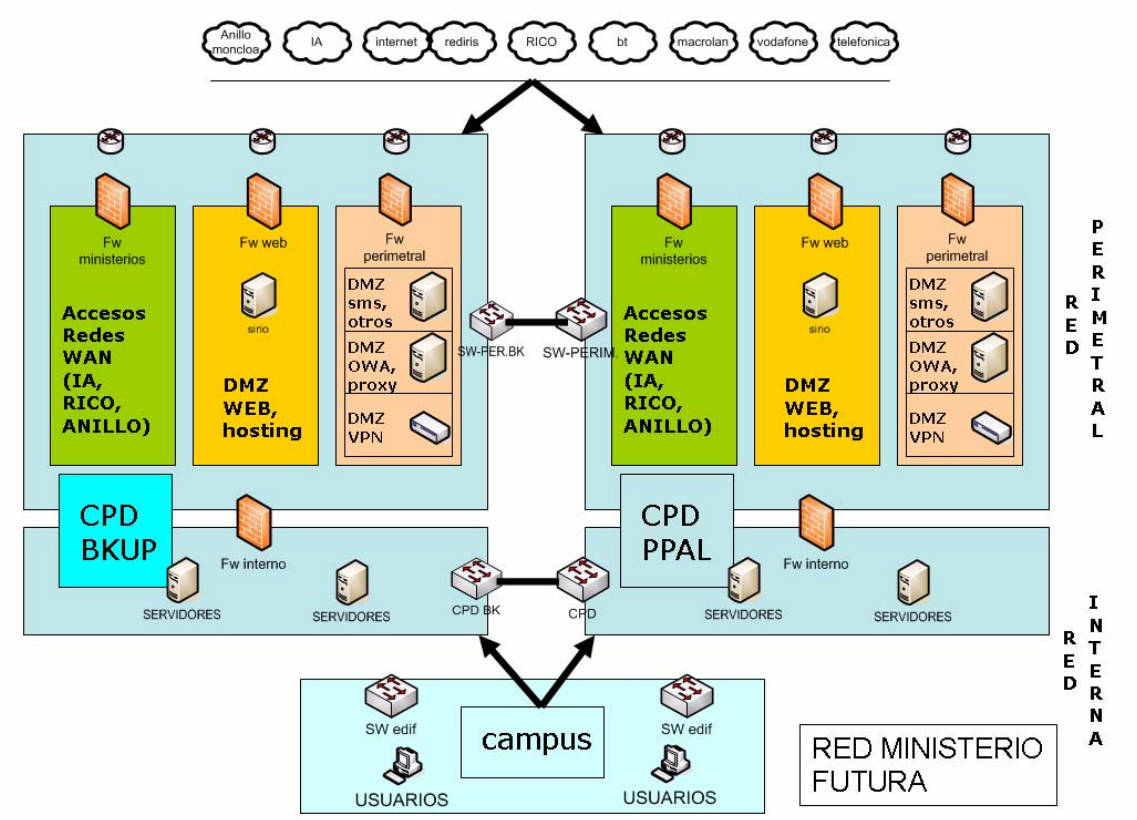
La filosofía de este diseño se basa en lo que se conoce como “extensión de VLANs”. El objetivo es tener en el Centro de Respaldo el mismo direccionamiento que en el Centro Principal y que estas redes se vean unas a otras como si de la misma localización física se tratara. Esta filosofía, además de permitir una creación progresiva y nada traumática del Centro

de Respaldo también implica una simplicación considerable del esquema final y de la administración del equipamiento, ya que permite tener equipos con exactamente la misma configuración en un lado que en otro e incluso dos equipos que actúan y se gestionan como uno solo, como puede ser el caso de algunos firewalls configurados en Activo/Pasivo.

De igual forma, se busca simplificar la duplicación de la conectividad entre redes externas, como son las redes de otros Ministerios, Internet, el Anillo, BT, etc. Este diseño permite además poder tener múltiples fallos y que aún sea posible el tránsito del tráfico de forma que éste pase del Centro de Respaldo al Principal y viceversa de manera transparente al usuario.

La extensión de VLANs supone que NO es necesario modificar el esquema de enrutamiento existente hasta el momento, se trata simplemente de un desdoblamiento físico de la red, como si de un espejo se tratara.

Con todo esto, el planteamiento final, se recoge en el siguiente esquema de bloques:



4.3. Protocolos y soluciones técnicas utilizadas

A continuación se comentan las tecnologías o técnicas en lo que se basa el diseño de la redundancia del Centro de Respaldo (PIX, HSRPs/ VRRP, DNS, GLBP, otras…).

- Spanning Tree Protocol: Solución de nivel 2
- VRRP: En un entorno con rutas estáticas, un fallo en el equipo al cual apuntan dichas rutas, provocaría la caída de la red, teniendo que recuperar el estado de dicho equipo o cambiar la ruta estática de todos los equipos para que apunten a otro equipo de backup,

para que la red vuelva a su correcto funcionamiento. El Virtual Router Redundancy Protocol (VRRP) elimina el punto de fallo inherente en este tipo de entornos. Así, dos o más equipos pueden actuar como un uno, compartiendo una misma dirección IP y dirección MAC, de forma que uno permanece activo, atendiendo las peticiones realizadas a dicha dirección mientras que el resto de equipos, permanecen a la espera de un fallo (intercambiando mensajes para detectar cuando se produce) para asumir dichas direcciones como suyas. De esta forma un usuario o elemento de red que apunte con una ruta estática a dicha dirección IP, no percibirá la caída del equipo al que apunta, pues otro equipo de backup asumirá dicho papel, de forma transparente al usuario. El VRRP es un protocolo estándar y en funcionamiento es muy similar al protocolo propietario HRSP de las plataformas del fabricante Cisco Systems.

- GLBP: Este protocolo permite, al mismo tiempo que tener una alta disponibilidad de gateway de salida, la posibilidad de balancear la carga entre varios.
- Servicios Publicos con varios proveedores: Para solucionar este problema existen diferentes posiblidades que se pueden valorar:
    - Tecnología de balanceo de DNS: existen varios fabricantes
    - Múltiples entradas en DNS
- Seguridad: Todos los dispositivos de Seguridad aquí implicados tienen la posibilidad de soportar Alta Disponibilidad, aunque de diferentes maneras: los equipos CISCO pueden configurarse en Activo/Pasivo o en Activo/Activo. El protocolo es propietario de Cisco y exclusivo para PIXes / ASA. Otros equipos de firewall como NOKIA utilizan VRRP: protocolo estándar de alta disponibilidad Activo/Pasivo con direcciones IP virtuales. El Concentrador VPN: tiene dos posibilidades VRRP o independientes.

4.4. Fases implantación arquitectura

Como ya se ha indicado, gracias a la filosofía del diseño, su implantación es relativamente sencilla y poco o nada traumática, por lo que abordarla poco a poco es totalmente factible e incluso recomendable. Fases propuestas:

- Fase 1: Conmutación Central y Firewall interno
- Fase 2: Enrutamiento principal y resto de Firewalls
- Fase 3: Enlaces externos (Anillo, BT, Internet y Ministerios) y servicios públicos