Introducción a Condor

Adrián Santos Marrero

17 de septiembre de 2004

Índice

1. ¿Qué es Condor?

Condor es un proyecto de la Universidad de Wisconsin-Madison (UW-Madison). Está ideado para aprovechar al máximo la capacidad computacional de una red de ordenadores. Normalmente sólo disponemos de la potencia del ordenador que estamos usando para ejecutar nuestros trabajos, y si, por ejemplo, tuviéramos que lanzar 100 veces un mismo programa con distinta entrada, tendríamos que hacerlo secuencialmente con la consecuente pérdida de tiempo. Condor nos permite ejecutar nuestro trabajo en tantas máquinas como haya disponibles, por lo que, en el mejor de los casos, nuestro trabajo finalizará en el tiempo que tarda en ejecutarse el más lento de nuestros procesos.

Condor pone a nuestra disposición toda la capacidad de cálculo desaprovechada en nuestra red, de esta manera, los recursos disponibles se incrementan considerablemente.

Condor nos será útil siempre que necesitemos ejecutar un trabajo intenso, tanto computacionalmente como en el tiempo. Al aprovechar solamente recursos ociosos no influye en el uso cotidiano de los ordenadores (ver sec. 2).

Además, nos permite:

- Conocer el estado de nuestros trabajos en cada momento
- Implementar nuestras propias políticas de orden de ejecución
- Mantener un registro de la actividad de nuestros trabajos
- Añadir tolerancia a fallos a nuestros trabajos

2. ¿Cómo funciona?

Básicamente, enviamos un trabajo a Condor, este lo pone en una cola, lo ejecuta y finalmente nos avisa del resultado.

Vamos a verlo un poco más de cerca para intentar comprender como funciona:

- Normalmente usaremos Condor porque queremos ejecutar repetidas veces un programa (posiblemente con diferente entrada) o porque se requiere mucho tiempo para su finalización y, mientras tanto, necesitamos seguir usando nuestra máquina.
- Inicialmente nuestro trabajo no necesita ninguna modificación para ser enviado a Condor. Sin embargo, tenemos que escribir un archivo de descripción del envío (ver sección 3.1.1).

- Una vez enviado a Condor, podemos seguirle la pista a nuestro trabajo con el comando `condor_q` (ver sección 3.2) o mediante un registro de actividad (fichero `Log`).
- Condor realiza periódicamente búsqueda de trabajos nuevos e intenta casarlos con recursos disponibles. Si no hay disponibles, el trabajo se quedará a la espera del próximo ciclo.
- Una vez Condor ha encontrado una máquina capaz de ejecutar el trabajo pendiente, lo envía y empieza la ejecución. Pueden ocurrir varias cosas mientras se está ejecutando un trabajo:
  - Lo más deseable sería que finalizara con éxito. Si esto ocurriera se enviarían las salidas del trabajo a donde haya especificado el usuario y se mandaría un correo electrónico al mismo con un resumen de lo ocurrido.
  - En el caso de que la máquina deje de estar utilizable (porque ha vuelto el usuario o alguno de los motivos explicados más abajo) el proceso deberá abandonarla. Si se estaba ejecutando en el universo "standard", se realizaría una imagen del estado actual del proceso (*checkpoint*) (ver sec. 3.1.2) y se finalizaría su ejecución. En el resto de universos, simplemente se instará al trabajo a que finalize su ejecución (para ello se le envía la señal SIGTERM y si, pasado un cierto tiempo, no muere se le envía SIGKILL).
  - Otra posibilidad es que el propietario del trabajo haya decidido borrarlo de Condor (ver sección 3.3) con lo que finalizará su ejecución inmediatamente.

A la hora de enviar nuestro trabajo hemos de tomar algunas precauciones:

- Tenemos que elegir un "universo" adecuado: en la mayoría de los casos nos bastará con el universo "vanilla" (ver sec. 3.1.2).
- Nuestro trabajo ha de ser capaz de ejecutarse en un sistema de procesamiento por lotes:
  - Ha de ser capaz de ejecutarse en "background". No ha de solicitar información interactivamente.
  - Puede usar STDIN, STDOUT y STDERR, pero estos serán archivos en vez de los periféricos habituales (teclado y pantalla).
  - Ha de organizar sus archivos de datos. Por ejemplo, separados por ejecuciones.

Notar que Condor no influye en el uso cotidiano de nuestros ordenadores, ya que solo utilizará máquinas ociosas, o lo que es lo mismo, las que cumplan los siguientes puntos:

- No se está usando el ratón o teclado
- No se está usando la máquina remotamente
- No se está usando para ejecutar ningún otro trabajo.

3. ¿Cómo lo uso?

Condor está compuesto de varias aplicaciones que nos permiten:

- Enviar trabajos a Condor: *condor_submit*.
- Controlar el estado de nuestros trabajos: *condor_q*.
- Borrar un trabajo: *condor_rm*.
- Obtener información del estado de Condor: *condor_status*.

3.1. Enviando trabajos. *condor_submit*

Para realizar el envío tenemos que especificar algunas opciones para que Condor sea capaz de manejar adecuadamente nuestro trabajo. Estas opciones incluyen qué comando se va a ejecutar, cuantas veces y con qué entrada, donde irá la salida de cada comando, requisitos de la máquina donde se ha de ejecutar, etc. Esta información se especifica en un fichero de texto que llamaremos fichero de descripción del envío. La sintaxis a seguir la veremos a continuación.

3.1.1. Fichero de descripción del envío

Este archivo será la entrada al comando *condor_submit*. Un ejemplo de fichero de envío se puede ver en el siguiente ejemplo:

```
############################
#
# foo.submit
#
# Ejemplo 1: Archivo simple de descripción del envío.
#
```

```
############################

Executable   = foo
Universe     = vanilla
input        = test.data
output       = foo.out
error        = foo.err
Log          = foo.log
Queue
```

Una vez guardado este fichero, le indicamos a Condor que lo ejecute de la siguiente manera:

```
[adrians@trevina ~]$ condor_submit foo.submit
Submitting job(s).
Logging submit event(s).
1 job(s) submitted to cluster 3.
```

Veamos con más detalle el contenido del archivo:

Executable: Especificamos la ruta y el nombre del archivo ejecutable. En el ejemplo solo se ha especificado el nombre, por lo que se espera que `foo` y `foo.submit` estén en el mismo directorio.

Universe: Elegimos un universo, por defecto se usará el universo "vanilla". Ver sección 3.1.2.

input: Archivo desde donde se leerá la entrada por defecto (stdin). Si no se especifica, se utilizará el archivo `/dev/null`.

output: Archivo donde se escribirá la salida del comando (stdout). Si no se especifica, se utilizará el archivo `/dev/null`.

error: Archivo donde se escribirá la salida de error del comando (stderr). Si no se especifica, se utilizará el archivo `/dev/null`.

Log: Archivo donde Condor almacenará un histórico de lo que le ha ocurrido a nuestro trabajo e información como su código de salida, errores relacionados con Condor, etc.

Queue Indica que Condor va a ejecutar una vez este trabajo, podemos especificar un número (por ejemplo `Queue 10` o escribir varias veces `Queue` con lo que se ejecutará tantas veces como hayamos escrito). Podemos especificar opciones para cada ejecución, por ejemplo: podemos tener un fichero de entrada (`input`) para cada ejecución de nuestro trabajo.

Veamos ahora otro ejemplo, esta vez un poco más complicado:

```
############################
#
# complex.submit
#
# Ejemplo 2: Archivo de descripción del envío usando
# Requirements y Rank.
#
############################

Executable   = complex
Universe     = vanilla
Requirements = Memory >= 64 && OpSys == "Linux" && Arch == "INTEL"
Rank         = Memory
input        = data.$(Process)
output       = out.$(Process)
error        = err.$(Process)
Log          = complex.log
Queue 10
```

En este ejemplo introducimos algunas opciones nuevas:

Requirements: Especificamos los requisitos que se han de cumplir para que nuestro trabajo se ejecute. En el ejemplo obligamos a que la máquina candidata tenga un procesador INTEL o compatible, esté ejecutando Linux y, además, no permitimos que tenga menos de 64MB de memoria RAM. En el caso de que no especifiquemos explícitamente los requisitos sobre arquitectura y sistema operativo, Condor los creará automáticamente para que nuestros trabajos se ejecuten en máquinas con la misma arquitectura y el mismo sistema operativo que la máquina desde donde se envió el trabajo. Para ver los requisitos finales de nuestro trabajo (una vez enviado a Condor) podemos ejecutar `condor_q -l`, este comando nos mostrará información detallada de cada trabajo enviado.

Rank: Define un valor numérico que expresa preferencia, es decir, dadas todas las máquinas candidatas a ejecutar nuestro trabajo, Condor evalúa esta expresión en cada una de ellas y elegirá aquellas donde su valor sea mayor. En el ejemplo, preferimos aquellas máquinas que tengan mayor cantidad de memoria RAM.

Para obtener más información acerca del uso de `Requirements` y `Rank` vea la sección 2.5.2 del manual de Condor.

En el ejemplo 2 también hemos usado unos nombres de archivo un tanto especiales en las opciones `input`, `output` y `error`. El uso de la cadena "$(Process)" implica que allí donde aparezca será sustituido por el número del trabajo que se va a ejecutar, es decir, en el ejemplo se crean 10 trabajos (`Queue 10`) y cada uno de ellos tendrá como entrada el archivo data.0, data.1, ... dependiendo de que número de trabajo sea. Lo mismo ocurrirá con los archivos de salida (`output`) y salida de error (`error`).

3.1.2. Universos

Para Condor, un "universo" define un conjunto de características que marcarán el entorno de ejecución de nuestro trabajo. Por ejemplo, para trabajos en Java existe un universo "Java". Este, por ejemplo, permitirá capturar las excepciones de la máquina virtual de Java o solo ejecutará los trabajos en máquinas con la máquina virtual disponible.

Básicamente podemos elegir entre tres universos:

vanilla Es el universo por defecto y, además, es el menos restrictivo con los trabajos que se pueden enviar (acepta cualquier programa escrito en cualquier lenguaje). La parte negativa es que no permite "checkpointing" o llamadas al sistema remotas (ver universo "standard" a continuación).

standard Este universo soporta "checkpointing" y llamadas al sistema remotas. Hacer "checkpointing" de un programa significa guardar en disco su estado actual de ejecución antes de parar el proceso. Gracias al "checkpointing", un programa se puede parar (guardándose su estado en un fichero), y más adelante se puede volver a ejecutar desde el punto exacto en que se abortó. Para que un programa pueda ser enviado a este universo ha de estar enlazado con las librerías de Condor (compilado usando *condor_compile*). Presenta algunas restricciones en los trabajos que se pueden enviar.

java Este universo está destinado para el envío de trabajos escritos en Java.

Para información más detallada acerca de cada universo puede visitar la sección 2.4.1 del manual.

3.1.3. Sobre el acceso a los ficheros

El único universo que dispone de un sistema de llamadas al sistema remotas es el "standard", debido a esto, cuando use otro universo (por ejemplo el "vanilla") asegúrese de que los archivos de entrada y salida de sus

trabajos se escriban o lean en directorios compartidos por NFS (es decir, visibles desde todas las máquinas). Un buen ejemplo es su directorio home (`/home/user/...`) ya que desde cualquier máquina se puede acceder a él. Un ejemplo de un directorio no compartido sería `/tmp/`, si crea un directorio `/tmp/my_condor_job/`, este solo será visible desde su máquina, por lo que cuando su trabajo se empiece a ejecutar en otra máquina y vaya a abrir un archivo en ese directorio se encontrará que no existe y no podrá continuar (estos errores aparecerán en el `Log` del trabajo).

3.2. Estado de los trabajos enviados. *condor_q*

Podemos obtener información acerca de nuestros trabajos con el comando *condor_q*:

```
[adrians@trevina ~]$ condor_submit myjob.submit
Submitting job(s).
Logging submit event(s).
1 job(s) submitted to cluster 1.

[adrians@trevina ~]$ condor_q


-- Submitter: trevina.iac.es : <161.72.81.178:1085> : trevina.iac.es
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
   1.0   adrians         7/13 12:37   0+00:00:00 I  0   0.0  myprog Example.1.0

1 jobs; 1 idle, 0 running, 0 held
```

Por defecto, este comando nos mostrará información de los trabajos que hemos enviado desde la máquina donde se ejecuta, en el ejemplo sería "trevina". La información que aparece en la salida sería:

- El `ID` es el identificador del trabajo y está formado por dos números:
  - El número antes del punto representa el "cluster". Un "cluster" es el conjunto de trabajos creado en un envío. Podemos ver un ejemplo en la salida del comando "condor_submit".
  - El número después del punto representa el trabajo dentro del cluster, como en el ejemplo solo creamos uno, será el trabajo 0. Trabajos sucesivos en el mismo cluster se nombrarían como 1.1, 1.2, ...

- El usuario que envío los trabajos.
- La fecha del envío.
- El tiempo de ejecución, en el formato: Días+HH:MM:SS.
- El estado actual del trabajo. Algunos valores posibles son:

  I: No se está ejecutando porque aun no se le ha asignado ninguna máquina (IDLE).

  R: Ejecutándose actualmente (RUNNING).

  H: El trabajo no se está ejecutando por deseo del propietario (HOLD). Ver el comando condor_hold, condor_release o la sección 2.6.3 del manual de Condor.
- La prioridad del trabajo que ha especificado el usuario. Ver comando condor_prio o la sección 2.6.4 del manual de Condor.
- El tamaño de la imagen del trabajo en megabytes.
- Y por último, el nombre del ejecutable.

En el caso de que un trabajo no se esté ejecutando, este comando también nos permite conocer el motivo gracias a la opción `-analyze`. Por ejemplo:

```
[adrians@trevina ~]$ condor_submit myjob.submit
Submitting job(s).
Logging submit event(s).
1 job(s) submitted to cluster 1.

[adrians@trevina ~]$ condor_q -analyze


-- Submitter: trevina.iac.es : <161.72.81.178:39869> : trevina.iac.es
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
---
001.000:  Run analysis summary.  Of 187 machines,
    187 are rejected by your job's requirements
      0 reject your job because of their own requirements
      0 match, but are serving users with a better priority in the pool
      0 match, but prefer another specific job despite its worse user-priority
      0 match, but will not currently preempt their existing job
      0 are available to run your job
```

```
        No successful match recorded.
        Last failed match: Thu Sep 16 12:38:09 2004
        Reason for last match failure: no match found

WARNING:  Be advised:
   No resources matched request's constraints
   Check the Requirements expression below:

Requirements = ((Memory > 2147483647)) && (Arch == "INTEL") &&
(OpSys == "LINUX") && (Disk >= DiskUsage) &&
(TARGET.FileSystemDomain == MY.FileSystemDomain)
```

En el ejemplo podemos ver como el trabajo 1.0 tiene problemas para ejecutarse: nuestros requisitos (`Requirements`) han desechado 187 máquinas de las 187 candidatas. Además, *condor_q* nos sugiere que revisemos dicha expresión y nos la muestra en su salida (en el ejemplo vemos como el límite mínimo de memoria RAM es excesivo).

Para más información puedes visitar la página del manual de *condor_q*, la sección 2.6.1 o la sección 2.6.5 del manual de Condor.

3.3. Borrando trabajos. *condor_rm*

Usaremos *condor_rm* para borrar un trabajo de la cola de Condor:

```
[adrians@trevina ~]$ condor_q


-- Submitter: trevina.iac.es : <161.72.81.178:1085> : trevina.iac.es
 ID      OWNER            SUBMITTED     RUN_TIME ST PRI SIZE CMD
   2.0   adrians         7/13 12:46   0+00:00:01 R  0   0.0  myprog Example.2.0

1 jobs; 0 idle, 1 running, 0 held

[adrians@trevina ~]$ condor_rm 2.0
Job 2.0 marked for removal
```

Podemos especificar tanto un trabajo como un cluster, en el ejemplo anterior, si hubiésemos ejecutado `condor_rm 2` habríamos borrado todos los trabajos del cluster 2.

Notar que no podemos borrar trabajos que no nos pertenezcan.

Para más información puede visitar la página del manual o la sección 2.6.2 del manual de Condor.

3.4. Estado de Condor. *condor_status*

Este comando nos permitirá conocer el estado de Condor:

```
[adrians@trevina ~]$ condor_status

Name          OpSys       Arch   State      Activity   LoadAv Mem   ActvtyTime

vm1@aciano.ia LINUX       INTEL  Claimed    Busy       1.030   501  0+16:56:02
vm2@aciano.ia LINUX       INTEL  Claimed    Busy       0.990   501  0+00:59:48
agracejo.iac. LINUX       INTEL  Claimed    Busy       1.030   500  0+21:00:39
vm1@agrimonia LINUX       INTEL  Claimed    Busy       1.010  1826  0+00:09:36
vm2@agrimonia LINUX       INTEL  Claimed    Busy       1.000  1826  0+00:09:32
alfalfa.iac.e LINUX       INTEL  Owner      Idle       0.000   248  0+00:32:55
...
tonina.iac.es SOLARIS29   SUN4u  Claimed    Busy       1.000   128  0+21:56:24
toro.iac.es   SOLARIS29   SUN4u  Unclaimed  Idle       0.000   128  0+00:00:04
tuno.iac.es   SOLARIS29   SUN4u  Owner      Idle       0.040   640  0+01:33:15
vibora.iac.es SOLARIS29   SUN4u  Claimed    Busy       1.010   576  3+02:59:06
viola.iac.es  SOLARIS29   SUN4u  Claimed    Busy       1.010   256  0+01:40:35
zarza.iac.es  SOLARIS29   SUN4u  Claimed    Busy       0.660   256  0+00:01:06
zorro.ll.iac. SOLARIS29   SUN4u  Claimed    Busy       1.040   384  1+03:38:25

                     Machines Owner Claimed Unclaimed Matched Preempting

          INTEL/LINUX        75    33      41         1       0          0
        SUN4u/SOLARIS29      87    21      64         2       0          0

                Total       162    54     105         3       0          0
```

Este comando muestra información sobre cada una de las máquinas que forman el "pool" de Condor y un resumen del estado actual. En este resumen podemos comprobar, en cifras, el uso que se le están dando a las máquinas. Así, por ejemplo, podremos comprobar cuantas máquinas quedan disponibles para ejecutar trabajos mirando la columna "Unclaimed".

Para más información puedes visitar la página del manual de este comando.

4. Limitaciones

- Si su trabajo realiza muchas operaciones de entrada/salida (E/S) tenga en cuenta la sobrecarga que esto conlleva, probablemente no obtenga

una mejora muy grande enviándolo a Condor. Note que todas las operaciones de lectura/escritura sobre archivos se realizan sobre la red[1] por lo que sus trabajos se verán ralentizados.

- Si envía un trabajo al universo "vanilla" contemple que cuando sea expulsado de una máquina perderá todo lo procesado hasta ese momento (a no ser que haya tomado precauciones por su cuenta). Si envía un trabajo que planea que vaya a tardar varios días en finalizar su ejecución probablemente no obtenga mejoría usando Condor, en estos casos plantéese el uso del universo "standard".

- Tenga en cuenta que cada trabajo que envíe generará un proceso "shadow" en la máquina desde donde se hace el envío. Este proceso se encarga de resolver algunas cuestiones relacionadas con su trabajo, realmente no consume demasiada CPU pero si realiza muchos envíos puede llegar a ralentizar su máquina.

5. ¿Y si tengo problemas?

Existe a su disposición un portal dedicado a Condor en el I.A.C., la dirección es: `http://goya/inves/SINFIN/Condor/`. También puede ponerse en contacto con los administradores de Condor en la dirección de correo condor@iac.es.

[1] Los archivos residen físicamente en el home de un usuario especial por lo que todas las peticiones se realizarán sobre NFS.