# ECPaaS Autoscaler User Guide

## Autoscaler說明與使用前置

### 使用版本

Kubernetes版本：1.22、1.23

Metrics Server：啟用

### 功能說明

使用者可以對ECPaaS中的部署(deployments)設定自動伸縮，可以根據需求設定兩種類型的自動伸縮，分別是水平自動伸縮(HorizontalPodAutoscaler，HPA)與垂直自動伸縮(VerticalPodAutoscaler，VPA)。

水平自動伸縮的作用是，當目標部署的所有副本(replica)的平均CPU或記憶體使用量超過設定的數值時，可以自動擴充副本的數量，以分攤副本的負擔。而當平均使用量下降到設定的數值以下時，可以自動減少副本的數量，節省資源。

垂直自動伸縮的作用不是增減副本數量，而是去調整部署的副本內部的容器資源設定，當副本使用的資源持續逼近上限時，可以自動調整容器資源設定以獲得更多的資源。或是當部署中的副本實際使用的資源低於設定的數值時，可以自動調整容器資源設定以釋放出多餘的資源。

### 安裝必要元件

HPA為Kubernetes的內建CRD，所以相關的元件皆已安裝在Kubernetes內。而VPA相關的元件與CRD則尚未内建到Kubernetes內，所以要使用VPA功能前，必須先使用以下指令安裝需要的元件。

git clone <https://github.com/kubernetes/autoscaler.git>

cd autoscaler/vertical-pod-autoscaler

### 切換到對應Kubernetes環境使用的版本，1.23版本的Kubernetes使用以下VPA版本

git checkout vertical-pod-autoscaler-0.11.0

### 執行安裝腳本

./hack/vpa-up.sh

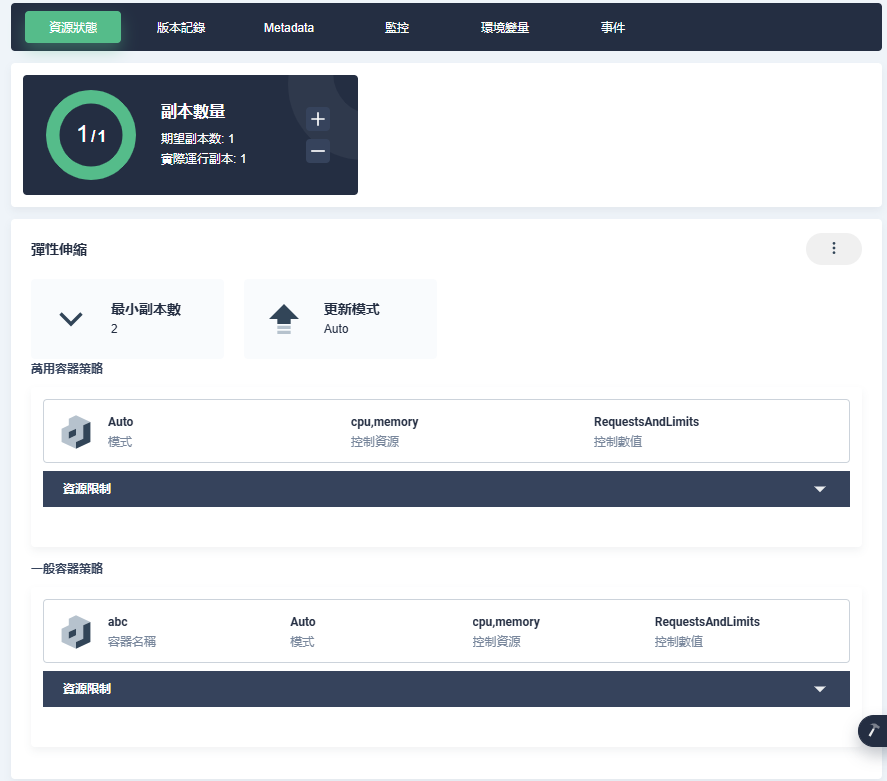
### 設定方式說明

在ECPaaS的「應用負載」的「工作負載」頁面中，進入「部署」分頁，找到想要設定自動伸縮的目標部署，自動伸縮只能設定在已經建立的部署上，並且一個部署同時只能設定一種自動伸縮。

進入目標部署的詳細資料頁面，點擊「更多操作」，若是尚未設定任何自動伸縮的部署可以在清單中找到2個選項：「設置水平自動伸縮策略」及「設置垂直自動伸縮策略」。



使用者可根據需求選擇要設置水平或垂直自動伸縮，這兩種自動伸縮的詳細設定內容將在後續的章節中說明，在此先跳過。當完成自動伸縮的設定後，會繼續停留在部署的詳細資料頁面，在「資源狀態」分頁中的副本數量調整按鈕下方可以看到「彈性伸縮」區塊出現：



根據設定的自動伸縮類型，區塊中顯示的欄位會有所不同，上圖中顯示的是垂直自動伸縮的區塊。點擊彈性伸縮區塊右上角的「…」按鈕可以選擇編輯或是刪除自動伸縮設定：



設置完自動伸縮後，「更多操作」中的選項也會改變，原本可以選擇設置兩種自動伸縮的選項會替換成「編輯自動伸縮策略」：



如果想要設置另一種類型的自動伸縮，需要將原本的自動伸縮設定刪除後，再次建立新的自動伸縮設定，刪除方式目前只能從「彈性伸縮」區塊的「…」按鈕中選擇。若是直接刪除部署，自動伸縮設定也會一併被刪除，不需擔心相同名字的部署會被先前的自動伸縮設定影響。

離開詳細資料頁面到部署清單頁面，可以在設置了自動伸縮的部署上看到相對應的記號及說明：

未設置：

一張含有 文字, 螢幕擷取畫面, 字型, 標誌 的圖片

AI 產生的內容可能不正確。

水平自動伸縮：



垂直自動伸縮：



## 水平自動伸縮/HorizontalPodAutoscaler

### 說明

在水平自動伸縮中，可以設定目標部署的最小和最大副本數量，以及CPU使用率或記憶體使用量，CPU及記憶體的使用量可以同時設定，當兩種使用量同時設定時，只要有任一條件達成，就會增加副本；當兩種使用量都低於設定的目標值時，才會減少副本。

水平自動伸縮定中，也可以更細節地設定增加或減少副本時的行為(behavior)，增減副本的行為皆有預設值可直接套用，方便使用者快速創建水平自動伸縮設定。

### 創建方式及欄位說明

按照先前說明的方式，進入部署列表找到想要設置自動伸縮的目標部署，這裡以「test」部署為例，可由「狀態」欄位中是否有時鐘圖示確認是否有設置任何自動伸縮設定：



上圖中，「test」部署沒有設置任何自動伸縮設定，而「bbox」則有設置，將滑鼠游標移動至「bbox」的「狀態」欄位，文字後方的小時鐘圖案上就可以看到「bbox」實際上是設置了哪一種類的自動伸縮。

確認目標部署尚未被配置任何自動伸縮設定後，點擊進入詳細資料頁面：



點擊「更多操作」，在出現的列表中找到「設置水平自動伸縮策略」：



點擊「設置水平自動伸縮策略」並在跳出的頁面中輸入所需資訊創建水平自動伸縮：



第一頁是水平自動伸縮的基本資訊，以下是欄位介紹：

**資源名稱**：此欄位會自動使用目標部署的名稱，用來確認新增的水平自動伸縮是否會套用到正確的目標部署上。

**CPU目標使用率**：當部署中的所有副本的平均CPU使用率達到此欄位的設定值時，會自動增加目標部署的副本數量。單位是%，可以輸入的範圍是1 ~ 100之間的數字，包含頭尾。如果不想要根據CPU使用率來增減副本，可以將此欄位清空。

**記憶體目標使用量**：當部署中的所有副本的平均記憶體使用量達到此欄位的設定值時，會自動增加目標部署的副本數量。單位是MiB，可以輸入的範圍是0.01 ~ 10000，包含頭尾，可以輸入到小數點後2位，如果小數點後超過2位數，會被四捨五入到小數點後2位。如果不想要根據記憶體使用量來增減副本，可以將此欄位清空。

\***CPU目標使用率**和**記憶體目標使用量**兩個欄位至少要設定一個；兩者同時設定時，只要有其中一個條件達標，就會觸發水平自動伸縮增加副本，當兩個條件皆未達標時，才會觸發水平自動伸縮降低副本數量。預設值是將CPU目標使用率設定在80%，記憶體目標使用量不設定。

**最小副本數**：設定目標部署中，可以由水平自動伸縮調整的最低副本數量，可輸入的範圍：1 ~ 2147483647，包含頭尾。預設值是1。

**最大副本數**：設定目標部署中，可以由水平自動伸縮調整的最高副本數量，此欄位的數值不可低於最小副本數欄位中的值，可以相等。可輸入的範圍：最小副本數 ~ 2147483647，包含頭尾。預設值是1，表示目標部署的副本數量不會超過1，此欄位在設定時會需要使用者自行調整至預設值以外的數量，否則水平自動伸縮等於沒有作用。



第二頁是增加副本行為的設定頁面，以下是欄位介紹：

**穩定窗口秒數**：根據此欄位設定的時間，觀察時間內的使用資源變化量來決定是否要增加副本。單位為秒，可輸入範圍：0 ~ 3600，包含頭尾。設定的時間越長，就需要更多時間來決定是否要增加副本，但同時也能避免副本數量增加過於快速。預設值為0秒。

**選擇策略**：選擇策略與策略設定是彼此關聯的設定，在策略設定中可以設定多組策略，而選擇策略則是用來選擇「使用哪一個策略」的設定。可以選擇的選項有：Max、Min、Disabled。Max選項會選擇在策略設定中影響最大的那個；Min選項則是選擇影響最小的那個。如果策略設定中僅有一組策略存在，那Max和Min選項就沒有差別。Disabled選項可以完全關閉增加副本行為，即便設定的目標CPU或記憶體使用量達標，也不會增加副本。預設值為Max。

**策略設定**：在策略設定中可以設定不同類型的策略，每一種類型的策略只能存在一組，目前有「數量」及「百分比」兩種類型，並且在策略設定中最少需要有一組策略存在，不限類型。除了類型，每一組策略還需要設定「數值」及「週期秒數」2個欄位，三個欄位缺一不可。每一個策略代表在設定的週期秒數內，可以增加的固定副本數量或是副本數量百分比，例如上圖中的例子，這兩個策略是增加副本行為的預設策略，分別代表「在15秒內增加4個副本」，以及「15秒內增加100%的副本數量」，會根據選擇策略欄位的選項來決定當下是哪一個策略會生效：**在上圖的例子中**，如果目標部署已經有超過4個副本存在，那百分比策略的影響就會大於數量策略，此時如果選擇策略是Max，那採用的就會是百分比策略；如果選擇策略是Min，那就會是數量策略生效。



第三頁是減少副本行為的設定頁面，可以設定的欄位與增加副本行為頁面的欄位相同，僅有部分欄位的預設值不同。以下是欄位介紹：

**穩定窗口秒數**：根據此欄位設定的時間，觀察時間內的使用資源變化量來決定是否要減少副本。單位為秒，可輸入範圍：0 ~ 3600，包含頭尾。設定的時間越長，就需要更多時間來決定是否要減少副本，但同時也能避免副本數量減少過於快速。預設值為300秒。

**選擇策略**：選擇策略與策略設定是彼此關聯的設定，在策略設定中可以設定多組策略，而選擇策略則是用來選擇「使用哪一個策略」的設定。可以選擇的選項有：Max、Min、Disabled。Max選項會選擇在策略設定中影響最大的那個；Min選項則是選擇影響最小的那個。如果策略設定中僅有一組策略存在，那Max和Min選項就沒有差別。Disabled選項可以完全關閉減少副本行為，即便設定的目標CPU或記憶體使用量都低於設定的目標，也不會減少副本。預設值為Max。

**策略設定**：在策略設定中可以設定不同類型的策略，每一種類型的策略只能存在一組，目前有「數量」及「百分比」兩種類型，並且在策略設定中最少需要有一組策略存在，不限類型。除了類型，每一組策略還需要設定「數值」及「週期秒數」2個欄位，三個欄位缺一不可。每一個策略代表在設定的週期秒數內，可以減少的固定副本數量或是副本數量百分比，例如上圖中的例子，這個策略是減少副本行為的預設策略，代表「15秒內減少100%的副本數量」。若有多個策略，會根據選擇策略欄位的選項來決定當下是哪一個策略會生效。

增加與減少副本行為中，所有欄位都有預設值可以使用，如果使用者不想要這麼細節的設定，可以直接點擊下一步跳過直到創建按鈕出現。只有第一頁的基本資訊頁面會需要使用者調整一下最小和最大副本數，否則即便水平自動伸縮生效，也沒辦法將副本數增加或減少。

### 修改方式

在創建了水平自動伸縮以後，如果需要修改設定，一樣必須先進到目標部署的詳細資料頁面，點擊「更多操作」，找到選單中的「編輯自動伸縮策略」，點選後會彈出編輯頁面讓使用者設定：





編輯頁面的第一頁一樣是基本資訊頁面，所有欄位都可以編輯，欄位介紹可以參考創建方式中的說明。編輯時僅有一點需要注意：**CPU目標使用率**和**記憶體目標使用量**這兩個欄位，如果在創建時有欄位是不設定，也就是清空的狀況下，在此編輯頁面會顯示為0，例如上圖中的記憶體目標使用量，如果要維持不設定，需要使用者將0清除。原本有設定的欄位，如上圖中的CPU目標使用率，要改為不設定的話，需要將原本的值清空。如果是要設定新的值，直接將原本欄位中的數值改為新的值即可。和創建時相同，CPU或記憶體兩個欄位至少要設定其中一個。

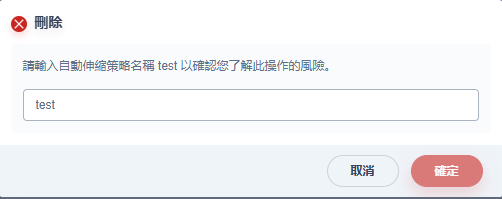
第二頁及第三頁的增減副本行為設定都和創建時的設定方式相同。

### 刪除方式

刪除水平自動伸縮的方式為，進入目標部署的詳細資料頁面的「資源狀態」分頁，在「彈性伸縮」區塊的右上角，點擊「…」按鈕，並找到刪除：



點擊刪除後會彈出確認視窗，輸入完需要的內容後點擊紅色確定按鈕完成刪除：



## 垂直自動伸縮/VerticalPodAutoscaler

### 說明

在垂直自動伸縮中，自動調整的不是副本數量而是副本使用的資源大小，部署的副本當中的容器會有各自獨立的資源(CPU和記憶體)需求與限制，在沒有設置垂直自動伸縮的情況下，一開始設定的資源需求和限制就不會改變了，無法取得足夠資源的情況下會導致容器執行效率下降，甚至出現錯誤。垂直自動伸縮可以自動調整目標部署中各容器的資源需求與限制，根據實際使用情況來增減分配到的資源。

使用者可以在垂直自動伸縮設定中，針對目標部署當中的每個容器設定不同的資源設定，包括種類及上下限等等，或是僅設定一個可以套用到所有容器的萬用設定。

### 創建方式及欄位說明

按照先前說明的方式，進入部署列表找到想要設置自動伸縮的目標部署，這裡以「test」部署為例，可由「狀態」欄位中是否有時鐘圖示確認是否有設置任何自動伸縮設定：



上圖中，「test」部署沒有設置任何自動伸縮設定，而「bbox」則有設置，將滑鼠游標移動至「bbox」的「狀態」欄位，文字後方的小時鐘圖案上就可以看到「bbox」實際上是設置了哪一種類的自動伸縮。

確認目標部署尚未被配置任何自動伸縮設定後，點擊進入詳細資料頁面：



點擊「更多操作」，在出現的列表中找到「設置垂直自動伸縮策略」：

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

AI 產生的內容可能不正確。

點擊「設置垂直自動伸縮策略」並在跳出的頁面中輸入所需資訊創建垂直自動伸縮：

一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

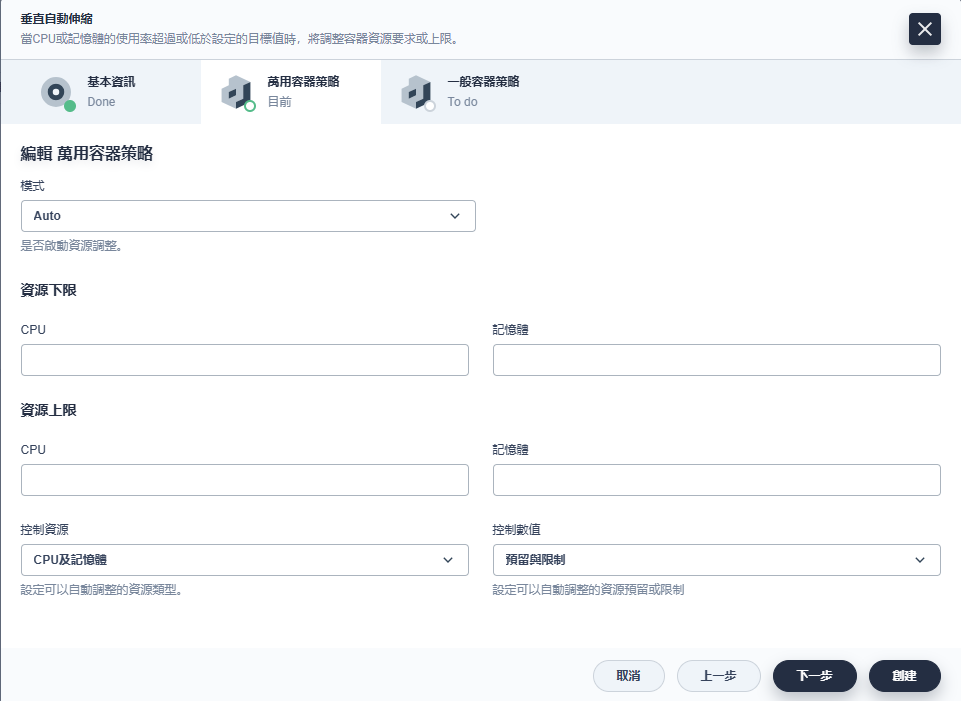
AI 產生的內容可能不正確。

第一頁是垂直自動伸縮的基本資訊頁面，以下是欄位介紹：

**資源名稱**：此欄位會自動使用目標部署的名稱，用來確認新增的垂直自動伸縮是否會套用到正確的目標部署上。

**更新模式**：此欄位要選擇垂直自動伸縮的更新模式，當垂直自動伸縮分析出目標部署中的容器可以使用多少資源時，會依照此欄位的設定來更新。一共有：Auto、Initial、Off三種模式。Initial模式下，垂直自動伸縮只會調整「新建」的副本中的容器資源設定。Auto模式則是除了Initial模式的功能，還會重啟已經存在的副本，以調整其中的容器資源設定。Off模式則是可以關閉調整功能，可以用在需要暫時停止垂直自動伸縮功能，但又不想刪除整個設定的情況。預設值為Auto模式。

**最小副本數**：垂直自動伸縮進行調整容器資源設定時，所需要存在的最低副本數量要求。可設定範圍：1 ~ 2147483647，包含頭尾。預設值為2。當設定為1的情況下，只要有1個副本存在，垂直自動伸縮就會嘗試去調整容器資源需求，如果更新模式又是Auto模式，就有可能讓唯一的副本被重啟，進而導致服務中斷，所以建議設置為2以上的數字。



第二頁是萬用容器策略設定頁面，萬用容器策略就是可以套用到所有容器的設定，此設定是必須設置的，當中的所有欄位皆有預設值，若使用者不想調整其中的設定，可以點擊「下一步」跳過或是點擊「創建」按鈕完成垂直自動伸縮設定。以下是欄位介紹：

**模式**：在此欄位可以控制垂直自動伸縮是否可以調整容器中的資源需求。一共有：Auto、Off兩種模式。Auto模式會開啟垂直自動伸縮調整資源的能力；Off模式則是關閉調整資源的能力。預設為Auto模式。

**資源下限**：資源下限又分成CPU和記憶體兩個欄位，垂直自動伸縮調整資源時，無法調整至低於此欄位的數值。

* CPU欄位，單位是**core**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置下限。
* 記憶體欄位，單位是**MiB**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置下限。

**資源上限**：資源上限又分成CPU和記憶體兩個欄位，垂直自動伸縮調整資源時，無法調整至高於此欄位的數值。

* CPU欄位，單位是**core**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置上限。
* 記憶體欄位，單位是**MiB**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置上限。

**控制資源**：設定垂直自動伸縮可以調整的資源類型，一共有：CPU及記憶體、CPU、記憶體，等三種選項。預設為「CPU及記憶體」，代表垂直自動伸縮可以調整CPU以及記憶體兩種資源。

**控制數值**：設定垂直自動伸縮可以調整的資源數值，一共有：預留與限制、只有預留，等兩種選項。預設為「預留與限制」，代表垂直自動伸縮可以調整容器中資源的預留(request)與限制(limit)等數值。

第三頁為目標部署中，個別容器專屬的一般容器策略設定頁面，使用者可以為目標部署中的容器設定專屬的容器設定，在此設定的一般容器策略只能套用到對應的容器上，並且會覆蓋掉萬用容器策略中的設定，一般容器策略為選填項目，使用者如不想設置一般容器策略，可以直接點擊「創建」按鈕完成設定。



點擊上圖中的「添加一般容器策略」方塊，會進入一般容器策略設定頁面：



新增一般容器策略時，需要選擇對應的容器名稱，每個容器名稱只能有一個一般容器策略存在，除了容器名稱外的欄位皆與萬用容器策略中相同。以下是欄位介紹：

**容器名稱**：可以在下拉選單中選擇容器名稱，一般容器策略只會套用到此欄位選擇的容器上，每個容器也只能設定一個一般容器策略，。

**模式**：在此欄位可以控制垂直自動伸縮是否可以調整容器中的資源需求。一共有：Auto、Off兩種模式。Auto模式會開啟垂直自動伸縮調整資源的能力；Off模式則是關閉調整資源的能力。預設為Auto模式。

**資源下限**：資源下限又分成CPU和記憶體兩個欄位，垂直自動伸縮調整資源時，無法調整至低於此欄位數值。

* CPU欄位，單位是**core**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置下限。
* 記憶體欄位，單位是**MiB**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置下限。

**資源上限**：資源上限又分成CPU和記憶體兩個欄位，垂直自動伸縮調整資源時，無法調整至高於此欄位數值。

* CPU欄位，單位是**core**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置上限。
* 記憶體欄位，單位是**MiB**，可輸入的範圍是0.01 ~ 10000，包含頭尾。超過小數點後2位數的數值會被四捨五入至小數點後2位。清空代表不設置上限。

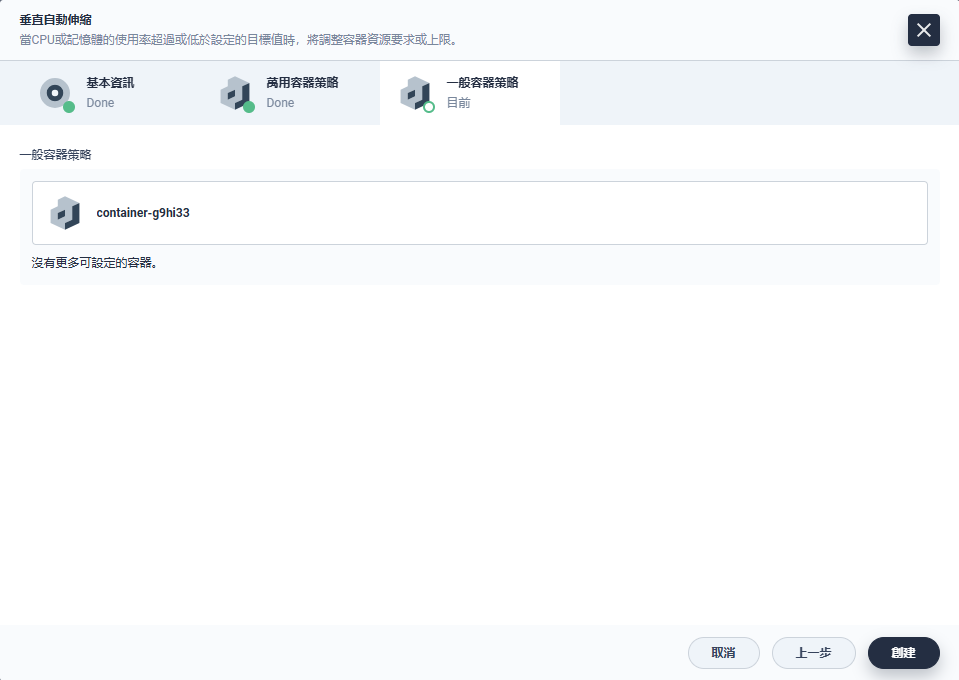
**控制資源**：設定垂直自動伸縮可以調整的資源類型，一共有：CPU及記憶體、CPU、記憶體，等三種選項。預設為「CPU及記憶體」，代表垂直自動伸縮可以調整CPU以及記憶體兩種資源。

**控制數值**：設定垂直自動伸縮可以調整的資源數值，一共有：預留與限制、只有預留，等兩種選項。預設為「預留與限制」，代表垂直自動伸縮可以調整容器中資源的預留(request)與限制(limit)等數值。

設定完成後，點擊右下角「√」保存設定，再點擊創建按鈕完成垂直自動伸縮的設定：

一張含有 螢幕擷取畫面, Rectangle, 行, 符號 的圖片

AI 產生的內容可能不正確。



### 修改方式

在創建了垂直自動伸縮以後，如果需要修改設定，一樣必須先進到目標部署的詳細資料頁面，點擊「更多操作」，找到選單中的「編輯自動伸縮策略」，點選後會彈出編輯頁面讓使用者設定：





垂直自動伸縮的編輯頁面與創建時的頁面完全相同，編輯一般容器策略時，需先將游標移至容器名稱上，才會出現刪除及編輯的按鈕：



完成編輯後，點擊確定按鈕完成編輯。

### 刪除方式

刪除垂直自動伸縮的方式為，進入目標部署的詳細資料頁面的「資源狀態」分頁，在「彈性伸縮」區塊的右上角，點擊「…」按鈕，並找到刪除：



點擊刪除後會彈出確認視窗，輸入完需要的內容後點擊紅色確定按鈕完成刪除：

