

CWAGEPS GPU平行化加速成果 與效能評估

李尚恩 陳鎰龍 吳秉柔 黃崇惟 曾喜絃 劉邦彥 李明營 吳佳瑩 吳蔚琳 郭珮萱

中央氣象署海象氣候組

NVIDIA Corporation



115年第四十屆天氣分析與預報研討會
A3-6



Introduction

Model name: CWAGFS-TCo (global)

Resolution: ~28 km, 72 layers (TCo383L72)

Dy-core: Semi-Lagrangian+Semi-implicit (2 time level)

Grid system: Octahedral reduced Gaussian grid

格點數: 東西向1552格、南北向768格、垂直72層



(figure from
<https://confluence.ecmwf.int/display/FCST/Introducing+the+octahedral+reduced+Gaussian+grid>)

如何加速 (平行計算)

- 將一個大任務，拆成非常細微的小任務，小任務分配給多個計算核心同時執行，達到縮短計算時間(加速)的效果。
- 例如整張地球網格，拆成每個格點
- 例如整個迴圈，拆成每個循環
- MPI (32): 將網格拆成32等分，將小網格分派給計算核心循序執行。
- GPU: 將網格拆成每個格點，並分配給所有CUDA核心執行。

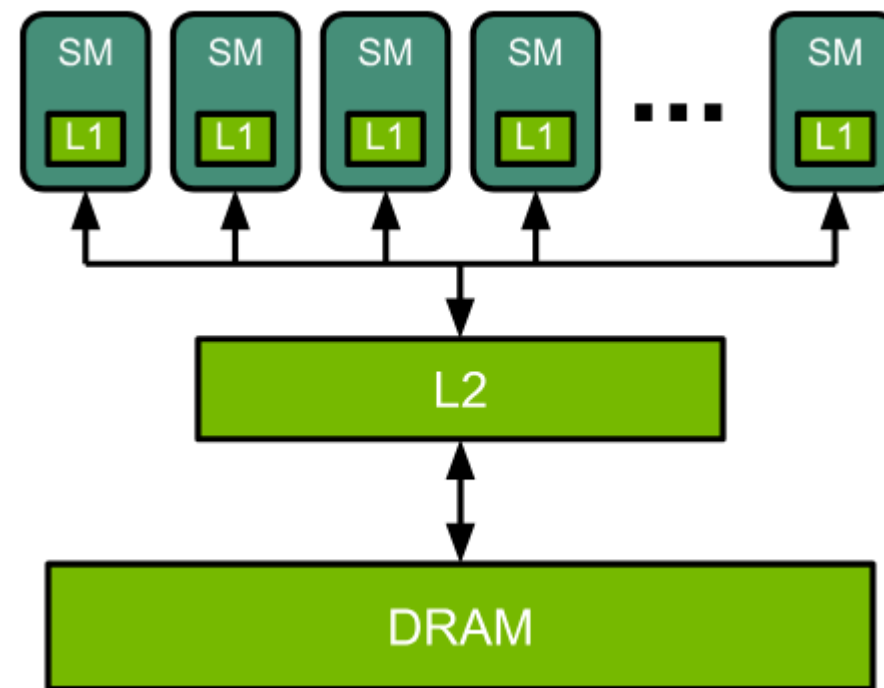
如何加速 (硬體)

- GPU含有多個流處理器(Streaming Processor, SM)
- 包含以下部分:
 - INT32
 - FP32
 - FP64
 - SFU
 - Tensor core
 - Register
 - Shared memory
- GPU平行策略:
- 將迴圈拆為每個循環(最小單位)，每個循環分配給CUDA Core執行。



如何加速 (硬體)

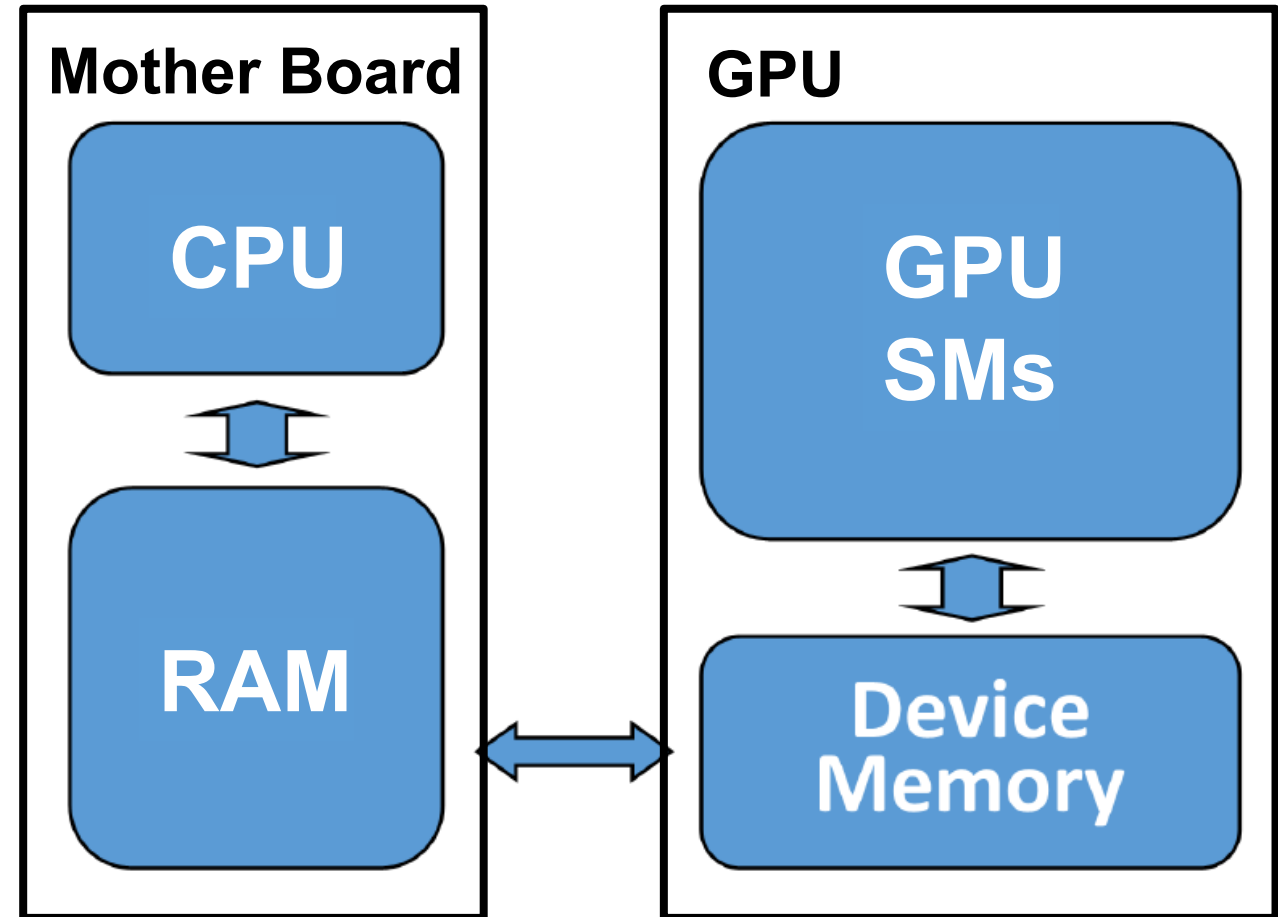
- 1張GPU有多個SM
- SM可以存取到GPU記憶體(DRAM)
(A100 有 80 GB)
- GPU記憶體為HBM記憶體
- (A100為HBM2e 1.94 TB/s)



<https://docscontent.nvidia.com/dita/00000186-1a08-d34f-a596-3f291b140000/deeplearning/performance/dl-performance-gpu-background/graphics/simple-gpu-arch.svg>

CPU、GPU互動方式

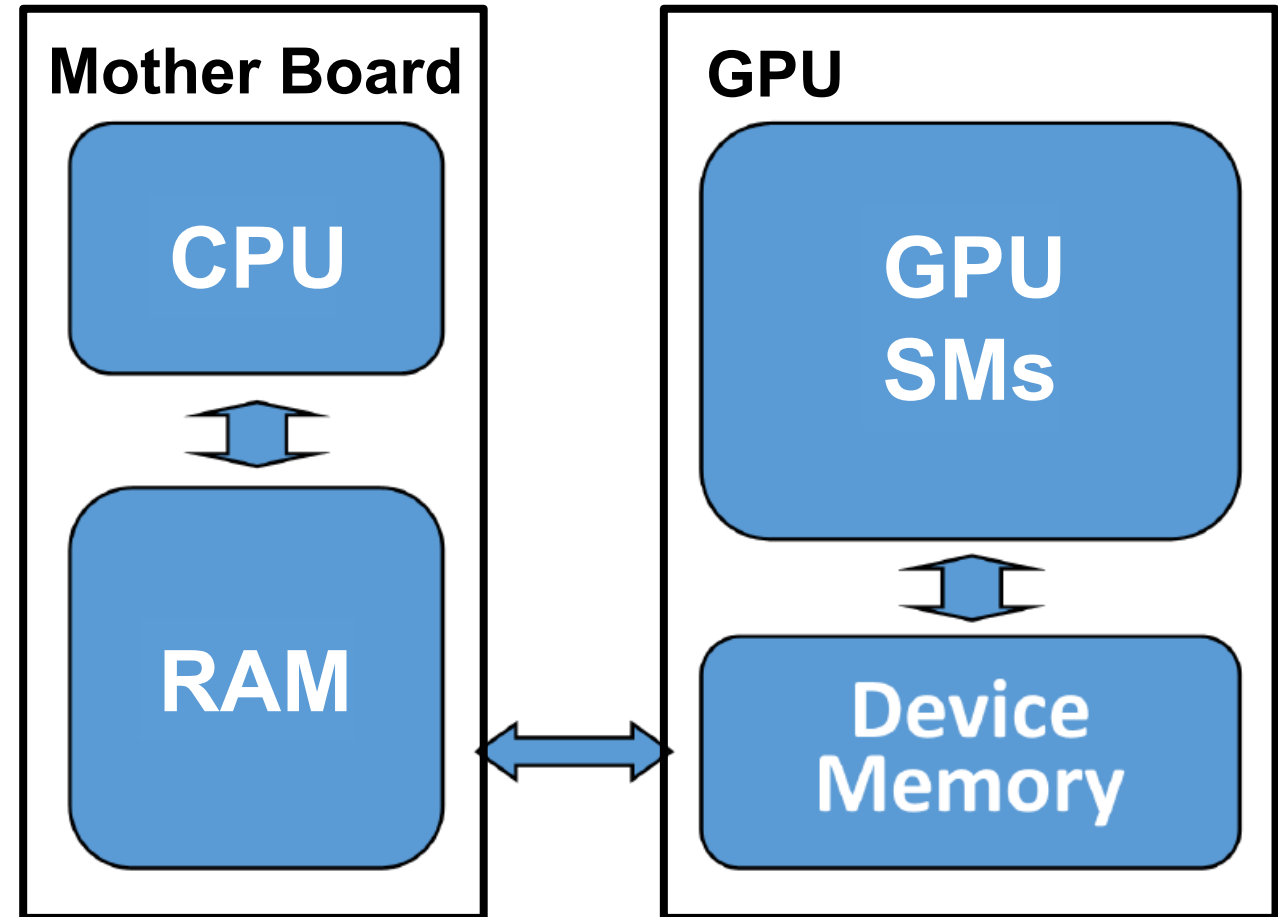
- CPU 計算 → 控制GPU
- DtoH、HtoD 頻寬:32 GB/s
- 計算前資料應先準備至GPU
- 減少資料進出GPU
- 資料流程
 - 積分開始前, 資料傳入GPU
 - 積分計算時, 無大陣列傳輸
 - 模式輸出時, 資料傳出GPU



(modified from https://www.openacc.org/sites/default/files/inline-files/OpenACC_Programming_Guide_0_0.pdf)

CPU、GPU互動方式

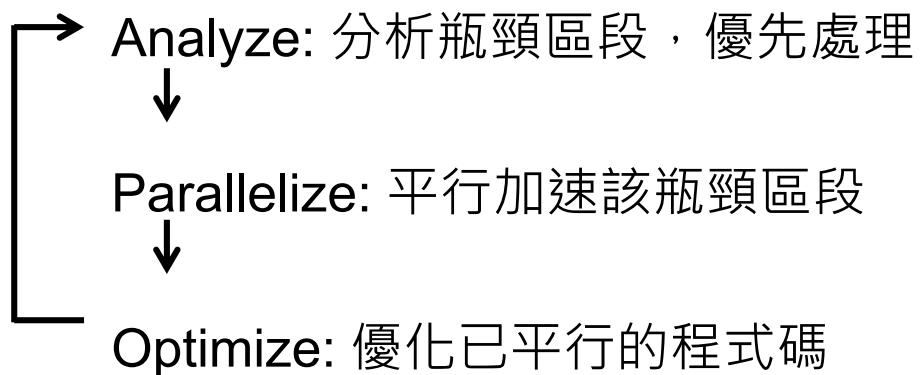
- CPU 計算 → 控制GPU
- DtoH、HtoD 頻寬:32 GB/s
- TCo的多卡平行計算策略:
 - 8個MPI各別對應8張A100
 - 全球網格切為8等分
 - 小網格再切為格點平行



(modified from https://www.openacc.org/sites/default/files/inline-files/OpenACC_Programming_Guide_0_0.pdf)

如何加速 (程式)

GPU程式開發流程:



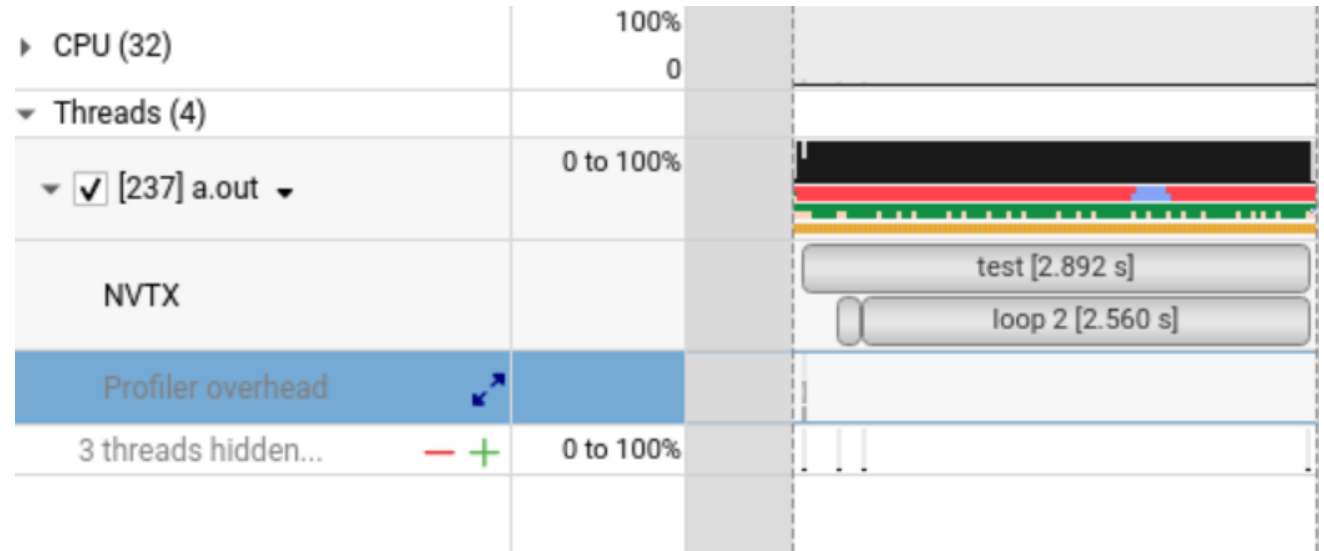
- 持續分析、聚焦在新的瓶頸、優化、持續提升效能。



Analyze

NVTX標註與Nsight分析:

```
call nvtxStartRange("test")
s = 2.
a = 1.
call nvtxStartRange("loop 1")
do i = 1, cnt
    a(i) = a(i) + i*s
end do
call nvtxEndRange
call nvtxStartRange("loop 2")
do i = 1, cnt
    v = i**0.25
    a(i) = a(i) + exp(1. + v)
end do
call nvtxEndRange
write(*,*) a(1), a(cnt)
call nvtxEndRange
```

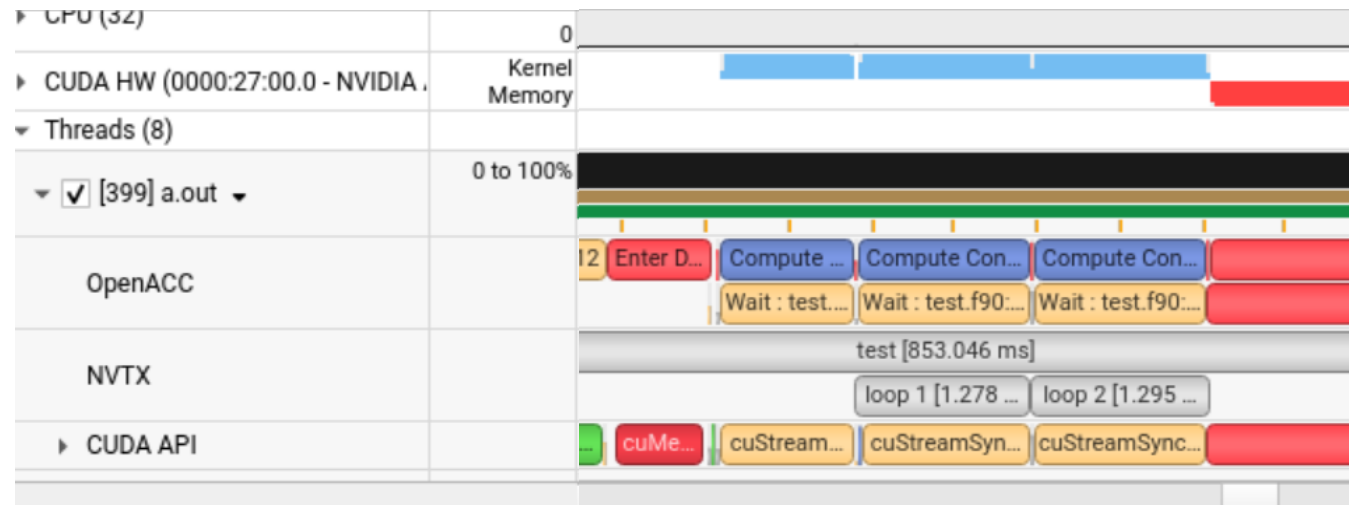


- NVIDIA Nsight Systems (效能分析工具)
- NVTX (CPU程式碼區段標示)
- CPU 計算時間: 2892 (ms)

Parallelize

平行網格內的所有格點:

```
s = 2.  
!$acc enter data create(a)  
!$acc parallel loop  
do i = 1, cnt  
    a(i) = 1.  
end do  
!$acc parallel loop  
do i = 1, cnt  
    a(i) = a(i) + i*s  
end do  
!$acc parallel loop  
do i = 1, cnt  
    v = i**0.25  
    a(i) = a(i) + exp(1. + v)  
end do  
!$acc exit data copyout(a)  
write(*,*) a(1), a(cnt)
```

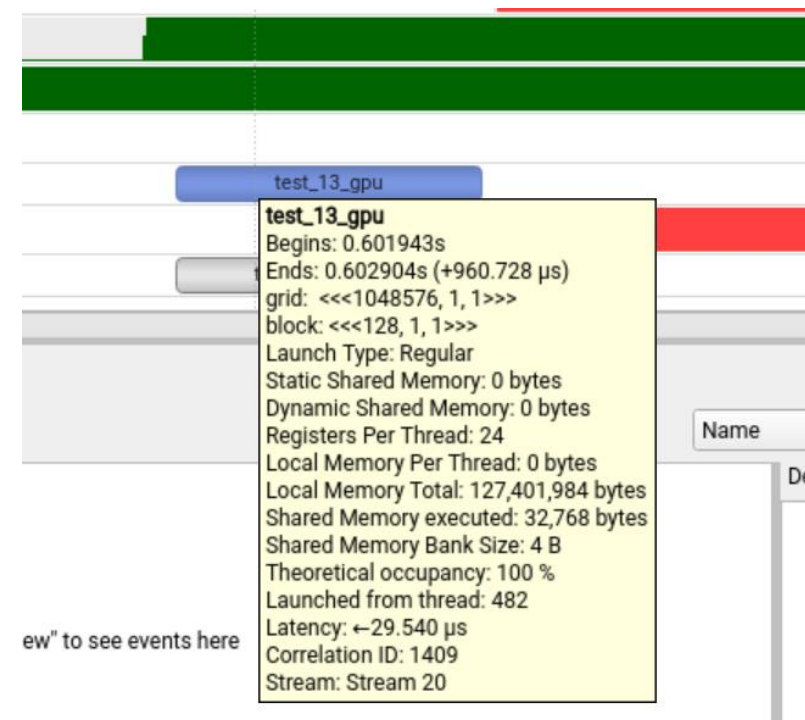


- OpenACC
- GPU 計算時間: 3.487 (ms)

Optimize

持續縮短GPU計算時間:

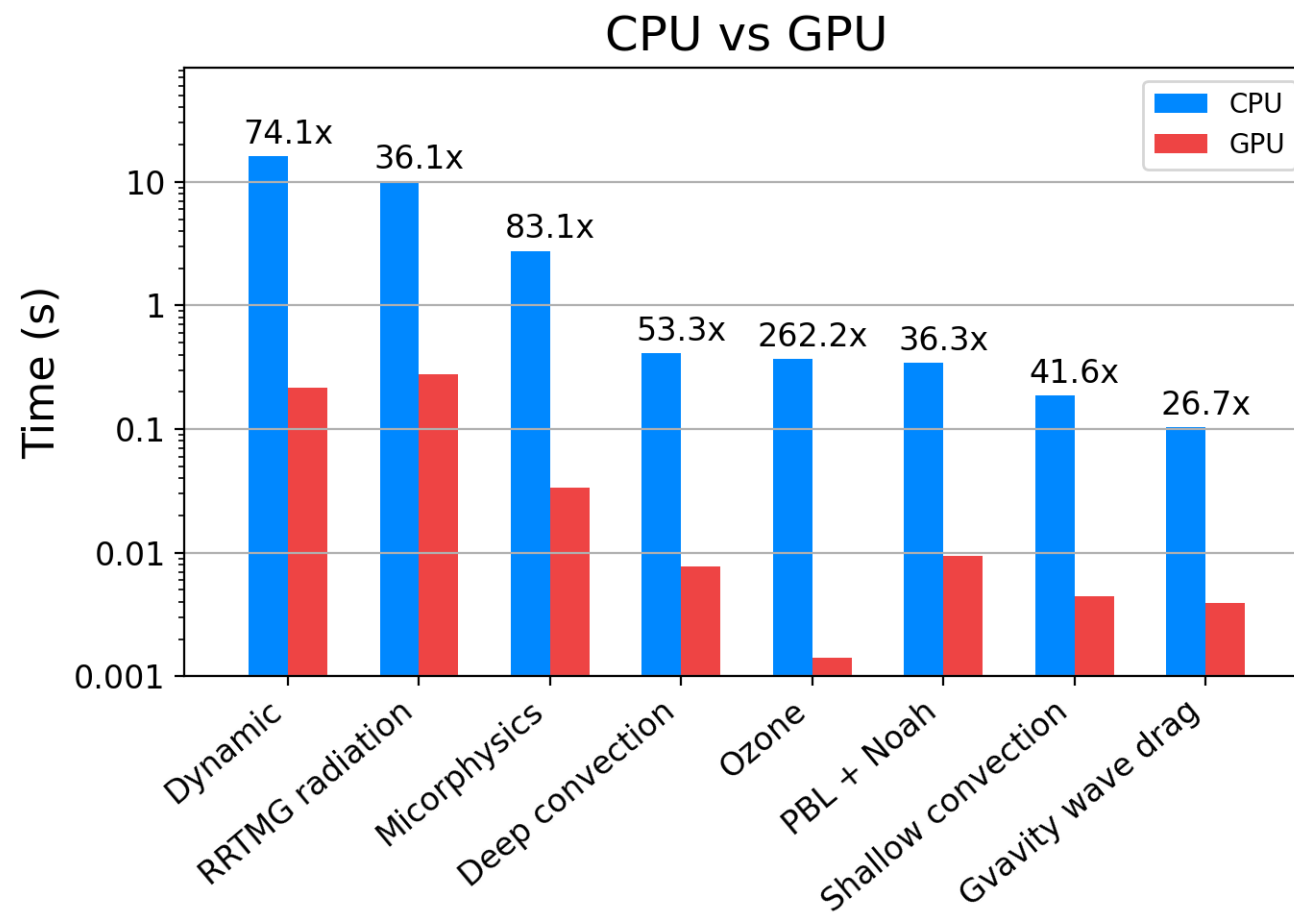
```
s = 2.  
!$acc enter data create(a)  
!$acc parallel loop private(v)  
do i = 1, cnt  
  a(i) = 1.  
  a(i) = a(i) + i*s  
  v = i**0.25  
  a(i) = a(i) + exp(1. + v)  
end do  
!$acc exit data copyout(a)  
write(*,*) a(1), a(cnt)
```



- 改為GPU友善的寫法
- GPU 計算時間: 0.96 (ms)

物理參數化加速成果

- 大多數參數化加速倍率約30~80倍。
- 雲微物理因可平行度高，計算量大，加速比達80倍以上。
- 積雲、淺對流部分可平行，加速比次之。
- 重力波拖曳、Noah計算量少，加速比較低。
- 臭氧修改演算法，加速比高。



CWAGEPS v3 GPU版效能評估

動 機

現階段CWAGFS-TCO大氣模式動力架構及物理參數化皆已完成GPU化。由於未來CWAGEPSv3可能以GPU形式上線，因此需做效能評估與能耗評估。

方 法

- 量測GPU、CPU之GEPSv3運算效能，積分至45天的執行時間。
- 測試時使用的配置：

代號	CPU	GPU	node
CPU496	496	--	10.33
GPU8	8	8	1
GPU4	4	4	0.5

- $$\text{node比} = \frac{\text{CPU時間}}{\text{GPU時間}} \times \frac{\text{CPU節點數}}{\text{GPU節點數}} \approx \text{GPU相對於CPU運算效益}$$

成果及規劃

- 初始化+積分計算+輸出
 - GPU8預報45天僅需**1.77小時**，可抵**73.5個CPU nodes**
 - GPU4預報45天僅需**2.88小時**，可抵**90.4個CPU nodes**，**效益較高**
- 能耗比量測：
 - 總能耗比**GPU僅約CPU的2成電量**

初始化+積分計算+輸出 (job總時間):

預報時長	1天	45天	node比
CPU496	16.8 分	12.6 時	--
GPU8	2.8 分	1.77 時	73.5 倍
GPU4	4.5 分	2.88 時	90.4 倍

實際量測能耗 (以pjstat -s直接測量) :

代號	總功耗(Wh)	能耗比
CPU496	18079	--
GPU8	3678.5	20.3%
GPU4	3264.4	18.1%

CWACFS TCo199/383L72 GPU版效能評估



動 機

- TCo預計以TCo199L72上線作業至CFSv3，因此需評估TCo199L72解析度之CPU版與GPU版效能差異。

方 法

- 量測GPU (CPU)版TCo199L72積分至480小時的平均所需時間，推算不同積分時長所需時間。
- 測試時使用的配置：

代號	CPU	GPU	node
CPU400	400	--	8.33
CPU6912	6912	--	144
GPU8	8	8	1
GPU2	2	2	0.25

- $$\text{node比} = \frac{\text{CPU時間}}{\text{GPU時間}} \times \frac{\text{CPU節點數}}{\text{GPU節點數}}$$

結 果

- 純積分計算，GPU8預報9960小時僅需**1.68小時**，可抵**114個CPU node**。
- GPU2所需時間僅為GPU8的**3倍**(使用卡數僅2張)。
- 相對於CPU總能耗比(直接測量): GPU8 **僅20.2%**，GPU2 僅22.6%。
- TCo383L72的GPU8較GPU6912快**75%**。

TCo199L72 純積分計算：

預報時長	1天 (平均)	1月 (平均)	14個月 (預估)	node比
CPU400	199 秒	99.6 分	22.96 時	--
GPU8	14.6 秒	7.28 分	1.68 時	114 倍
GPU2	44.6 秒	22.3 分	5.13 時	149 倍

TCo383L72 純積分計算：

預報時長	1步 (平均)	1天 (平均)	45天 (平均)
CPU6912	0.658 秒	3.95 秒	71 分
GPU8	0.367 秒	2.2 秒	40 分

(加入輸出，僅需48分)

整理



- GEPS 已使用OpenACC將積分計算移植至GPU。
- 使用1個A100 node, GEPS可與**70個**CPU node效能相當, TCo199L72可與**114個**CPU node效能相當。
- GPU總耗電量僅為CPU的**2成**。
- GPU利用數千個CUDA core進行平行計算, 以達減少計算時間的效果。

感謝聆聽