```
(env311) PS F:\test> & F:/test/env311/Scripts/python.exe f:/test/net/maximum performance.py
INFO:__main__: │ 初始化最大性能計算引擎...
INFO: __main__: ☑ 最大性能環境初始化完成
INFO:__main__: 主設備: gfx1010:xnack=
            計算單元: 18
INFO:__main__:
INFO:__main__: 最大工作組: 256
INFO:__main__:
            向量寬度: 1
INFO:__main__:
             並行隊列:8
INFO:__main__: ❷ 初始化高性能記憶體池...
INFO: __main__: ☑ 高性能記憶體池初始化完成
INFO:__main__: 總分配: 217.0 MB
INFO:__main__:
            池數量: 375
INFO:__main__: → 預編譯高性能kernel...
F:\test\env311\Lib\site-packages\pyopencl\cache.py:496: CompilerWarning: Non-empty compiler output
encountered. Set the environment variable PYOPENCL COMPILER OUTPUT=1 to see more.
 _create_built_program_from_source_cached(
INFO: main : ☑ 高性能kernel編譯完成
INFO: main : □ 啟動後台數據生成器...
INFO: main : ☑ 後台數據生成器啟動完成
INFO:__main__:
______
INFO:__main__: 则試設備: gfx1010:xnack-
INFO: __main__: 計算單元: 18
INFO: main :
            全局記憶體: 8176 MB
INFO: main :
             最大工作組: 256
INFO: main
🚀 單核最大性能測試:
INFO:__main__:
--- 測試大小: 1024 元素 (4.0 KB) ---
INFO:__main__: ● 最大性能測試 (大小: 1024, 迭代: 8)
INFO:__main__: ♠ 預熱GPU...
f:\test\net\maximum performance.py:490: RepeatedKernelRetrieval: Kernel 'maximum compute kernel' has been
retrieved more than once. Each retrieval creates a new, independent kernel, at possibly considerable
expense. To avoid the expense, reuse the retrieved kernel instance. To avoid this warning, use
cl.Kernel(prg, name).
 kernel = program.maximum_compute_kernel
INFO:__main__:☑ GPU預熱完成
INFO: main : 總時間: 0.190 ms
INFO:__main__:
            計算時間: 0.176 ms (92.6%)
INFO:__main__:
             記憶體時間: 0.013 ms (6.6%)
INFO:__main__:
            計算密度: 💧 優秀
INFO: __main__:
--- 測試大小: 4096 元素 (16.0 KB) ---
INFO:__main__: │ 最大性能測試 (大小: 4096, 迭代: 8)
INFO:__main__: 總時間: 0.199 ms
INFO: main : 計算時間: 0.158 ms (79.4%)
INFO: main :
            記憶體時間: 0.040 ms (19.9%)
INFO:__main__:
             計算密度: 🥚 優秀
INFO: main
--- 測試大小: 16384 元素 (64.0 KB) ---
INFO:__main__: │ 最大性能測試 (大小: 16384, 迭代: 8)
INFO: main :
            總時間: 0.551 ms
INFO: main :
            計算時間: 0.182 ms (33.0%)
INFO: main :
             記憶體時間: 0.363 ms (66.0%)
INFO: main :
             計算密度: 🔔 需優化
INFO: main :
--- 測試大小: 65536 元素 (256.0 KB) ---
INFO: main _: │ 最大性能測試 (大小: 65536, 迭代: 8)
INFO: __main__:
             總時間: 1.239 ms
INFO: main :
             計算時間: 0.278 ms (22.5%)
INFO: main :
             記憶體時間: 0.896 ms (72.4%)
```

計算密度: 🔔 需優化

INFO: main :

```
LLM Inference Optimization Techniques - Claude
2025/8/7 晚上10:57
 INFO:__main_
 --- 測試大小: 262144 元素 (1024.0 KB) ---
 INFO:__main__: │ 最大性能測試 (大小: 262144, 迭代: 8)
 INFO:__main__: 總時間: 6.341 ms
 INFO:__main__:
              計算時間: 0.491 ms (7.7%)
 INFO:__main__: 記憶體時間: 5.849 ms (92.2%)
 INFO:__main__:
               計算密度: 🔔 需優化
 INFO: main
  ● 大規模並行測試:
 INFO:__main__:
 --- 並行配置: 262144 元素, 4 塊 ---
 INFO: __main__: 💉 大規模並行測試 (總大小: 262144, 分塊: 4)
 INFO: main : 總時間: 4.681 ms
 INFO: main :
              平均塊時間: 0.762 ms
 INFO: __main__:
               最慢塊時間: 1.592 ms
 INFO:__main__: 並行效率: 0.65倍
 INFO: __main__:
              吞吐量: 56.00 M元素/秒
 INFO: main :
 --- 並行配置: 1048576 元素, 8 塊 ---
 INFO: main : 

√ 大規模並行測試 (總大小: 1048576, 分塊: 8)
 f:\test\net\maximum performance.py:511: RepeatedKernelRetrieval: Kernel 'streaming compute kernel' has
 been retrieved more than once. Each retrieval creates a new, independent kernel, at possibly considerable
 expense. To avoid the expense, reuse the retrieved kernel instance. To avoid this warning, use
 cl.Kernel(prg, name).
   streaming_kernel = program.streaming_compute_kernel
 INFO: main :
              總時間: 12.914 ms
 INFO: main :
              平均塊時間: 0.406 ms
              最慢塊時間: 0.480 ms
 INFO: __main__:
 INFO: main :
              並行效率: 0.25倍
 INFO:__main__:
               吞吐量: 81.19 M元素/秒
 INFO:__main__:
 --- 並行配置: 4194304 元素, 16 塊 ---
 INFO:__main__: 🚀 大規模並行測試 (總大小: 4194304, 分塊: 16)
 INFO: main : 總時間: 43.972 ms
 INFO:__main__:
              平均塊時間: 0.483 ms
 INFO: main :
              最慢塊時間: 0.636 ms
 INFO:__main__: 並行效率: 0.18倍
               吞吐量: 95.39 M元素/秒
 INFO: main :
 INFO:__main__:
 ______
 INFO:__main__:
 ■ 單核性能分析:
               1024 元素:
                              0.190ms, 計算占比: 92.6%
 INFO: __main__:
 INFO:__main__:
                  4096 元素:
                              0.199ms, 計算占比: 79.4%
 INFO: __main__:
                16384 元素:
                              0.551ms,計算占比: 33.0%
 INFO:__main__:
                              1.239ms, 計算占比: 22.5%
                65536 元素:
 INFO:__main__
                 262144 元素:
                              6.341ms, 計算占比:
                                              7.7%
 INFO: __main__
  累 最佳單核配置: 1024 元素 (計算占比: 92.6%)
 INFO:__main__:
 🚀 並行性能分析:
                 262144 元素 x 4 塊: 0.65倍效率, 56.00 MOPS
 INFO: main :
 INFO:__main__:
                1048576 元素 x 8 塊: 0.25倍效率, 81.19 MOPS
                4194304 元素 x 16 塊: 0.18倍效率, 95.39 MOPS
 INFO:__main__:
 INFO:__main_
  累 最佳並行配置: 4194304 元素, 16 塊 (吞吐量: 95.39 MOPS)
 INFO:__main__:
 INFO: main : ✓
               ▲ 極致性能!計算密度超過80%,已完全釋放APU潛力
 INFO: main :
  🦞 性能突破總結:
                ☑ 預編譯kernel + 激進編譯優化
 INFO: __main__:
```

https://claude.ai/chat/93e2dd20-5953-460b-bc82-b5ecc9a25d7c

INFO: main :

☑ 高性能記憶體池 + 零拷貝buffer

2025/8/7 晚上10:57