Conclusion

- 撰寫 GPU 程式需要對硬體有所了解
  - Warp divergence
  - Memory Bandwidth

- 通常 GPGPU 瓶頸在於記憶體頻寬

- 機器學習、AI 訓練資料量大

Conclusion

- 撰寫 GPU 程式需要對硬體有所了解
  - Warp divergence
  - Memory Bandwidth
- 通常 GPGPU 瓶頸在於記憶體頻寬
  - 機器學習、AI 訓練資料量大

Further Optimization Techniques

- CGMA ratio :

  https://www.sciencedirect.com/topics/computer-science/global-memory-access

- Bank conflict :

  https://blog.csdn.net/Bruce_0712/article/details/65447608

- CUDA streams:

  https://developer.download.nvidia.com/CUDA/training/StreamsAndConcurrencyWebinar.pdf