Developments in Computing Technology: GPUs

Mark Richardson, Technical Head, CEMAC

Welcome to first CEMAC seminar

- Will try to hold one every 3 months with some technical matter
- GPUs now becoming mainstream
- Potential future talks on
  - Hardware such as Intel XeonPhi (KNL) and ARM
  - Data management
  - Programming challenges
- Present case studies of projects undertaken

Agenda

UNIVERSITY OF LEEDS

- Introduction (MR 10 min)
- Case study (Sofya Titarenko 20 min)
  - Questions for Sofya
- Using GPUs on ARC3 (Mark Dixon 20 min)
  - Questions for Mark Dixon
- Round up (MR 5 min)

What is a GPU ?

UNIVERSITY OF LEEDS

- Originally for high speed graphic
  - Used to render 2D images from complex 3D scenarios
  - Many cores: earliest simplest was 16
  - latest >3000
- Since 2006 (NVIDIA GeForce 8800 GTX)
  - Even desktop cards have been **CUDA** capable
  - **Send** some intense numerical functions from the host processor to GPU cores
  - Consider the GPU as an “accelerator” using the OpenACC, OpenCL, OpenMP4.0 APIs
- Programming the GPU
  - CUDA (more recently can do **CUDA Fortran**)
  - Often re-write program sections with **C/C++** and add the CUDA constructs
  - OpenCL can be used with other (non-NVIDIA) accelerators
  - OpenMP 4.0 onwards has accelerator support for various target architectures

- Early versions only supported “single precision”
  - 32bit floating point representation
  - Extra work was needed for double precision
- Currently offering P100 Tesla “accelerators”
  - Can cope with double precision
  - 3584 “CUDA cores” in one card
  - Still requires PCI-e on a hosted Mother Board
  - Still requires codes to have CUDA API constructs

HPC resource

UNIVERSITY OF LEEDS

- For more than 25 years HPC systems have been:
  - Login node where users prepare the case for simulations
  - Compute nodes [convergence on Intel hardware due to costs]
  - Interconnect [+ method for transferring information (MPI)]
  - Data storage [probably needs a separate seminar talk]
  - Post processing node
- Trend was *"the more the merrier"* [See Top500 slides this talk]
- Recent development is towards systems using fewer Watts per FLOP
  - Leads to more complicated nodes
  - Hybrid with a mix of commodity Intel processors and GPUs
  - Also other Many-core technologies [Knights Landing : Xeon Phi]
  - Newer processor types [e.g. ARM64]
  - Green 500 list https://www.top500.org/green500/
- A GPU hybrid HPC resource will have "traditional" nodes with added GPU cards as many as 8 cards per node is possible.

Why now?

UNIVERSITY OF LEEDS

- More than 10 years on and the GPU platform has matured
  - Such that several major HPC facilities have GPU installed hardware
  - Can now host multiple (2,4,8 GPU cards) in a single node (usually 2xIntel Xeons)
    - i.e. Still require a host processor so “hybrid systems”
  - More recently some Fortran compilers added an API for CUDA “CUDA Fortran”
  - Can now work with double precision floating point representation
- More Scientific Software now GPU tuned
  - As with any parallel programming model you will rarely get an ideal speed-up of full runtime; but the tuned parts often can be.
  - Typically only small sections of a simulation will be treated in the accelerator regions
- UK systems available
  - ARC3 at Leeds (see talk from Mark Dixon)
  - JADE at Oxford is a Tier-2 national facility
    - Access through EPSRC RAP scheme
    - Not necessarily best configuration for CFD/Weather as targeting deep learning

Top 500 (1/2)

- TOP500 list (www.top500.org) is refreshed twice a year:
  - June in European ISC and Nov in Global Supercomputing Conference
  - Similar to "*Race to the Moon*" with prestige for those at top of league
  - Has been a tradition of "*More is better*"; now changing
- On the 2017 list the systems measured in units of PF (10^{15} FLOPS).
  - Essentially measures how well a computer system can process the LINPACK benchmark suite
  - Some other schemes are being developed (Graph500)
- For comparison a traditional desktop would register 25GFLOPS (10^{9} FLOPS).
- The range of speed in top 10 is from **10**PetaFLOPS to **95**PetaFLOPS.
  - (~$4x10^{6}$ faster than a single core desktop)
- **18** systems in the top 100 have NVIDIA hardware
- The top system recording **95**PF requires **15MW** power
- Third on the list is a system with hybrid nodes at CSCS Switzerland
  - also host to Meteo Swiss
  - A Cray system using Intel host and NVIDIA GPU
  - **20**PF sustained with **2.2**MW power consumption
  - It is also **10th** on green500 with registered FLOPS per Watt of **10.4**

Top500 (2/2)

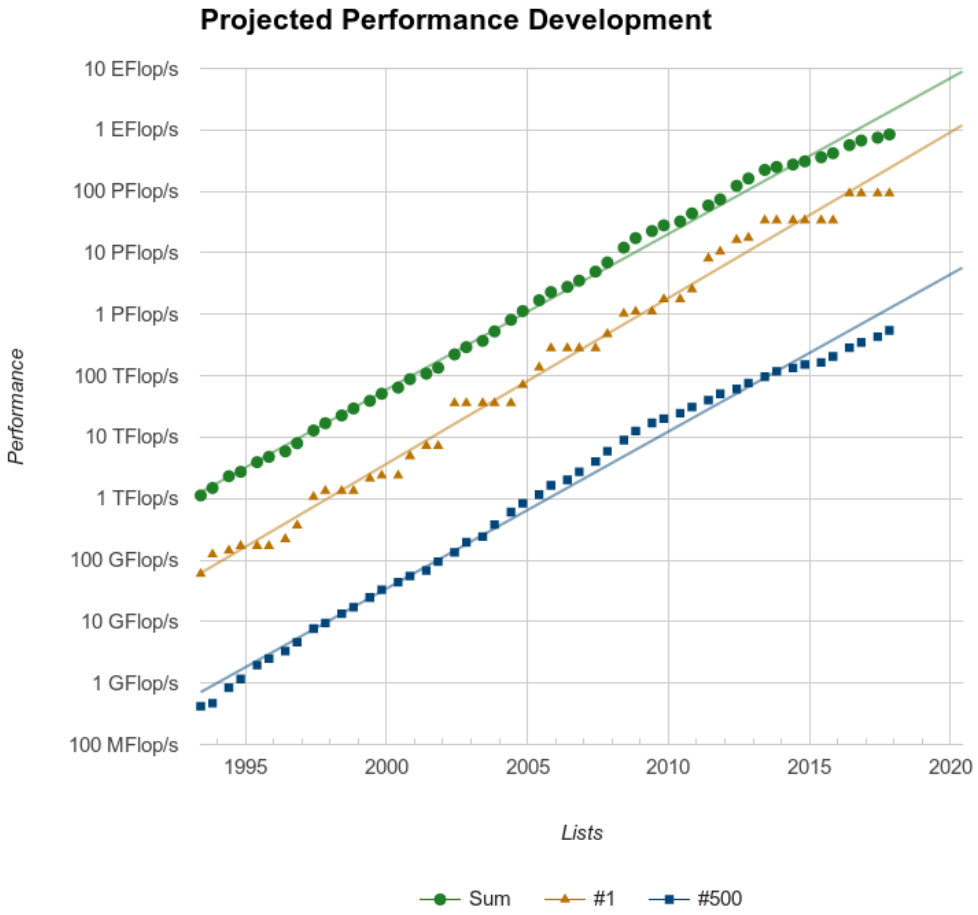
UNIVERSITY OF LEEDS

- "*In real world weather and climate codes have only ever been able to utilise between 10% and 15% of the system performance, hindered by striding through data, inter-node communications and I/O*." [from discussions at ECMWF HPC for NWP in Oct 2016]
- UK Archer system is currently #**79**
- UKMO **15**th on the list with **8**PF
  - (#46+#47 additional identical production systems)
- ECMWF #**27**+#**28** (identical production systems)
- **9** UK systems in top 100

Performance development of top500 list

UNIVERSITY OF LEEDS

Image from: https://www.top500.org/statistics/perfdevel/



Who is using GPUs?

- Molecular dynamics for N-body problem easiest concept.
- New discipline of "Deep learning", "Graph" and "map-reduce" analysis (there are top 500s for those too)
- Computations with large sections of "embarrassingly parallel" regions.
  - E.g. Smoothed particle hydrodynamics can see 5 codes listed as GPU capable (since 2004)
- Computational Fluid Dynamics of which weather and climate are a subset.
  - The weather codes enhanced according to NVIDIA:
    - COSMO
    - GEOS-5
    - HOMME CAM-SE
    - NEMO
    - NIM
    - WRF
- Computational Solid mechanics (see talk by Sofya)

Who is using GPU for weather and climate?

- Meteo-Swiss COSMO developmment
  - Closely associated with CSCS Piz Daint
  - Earliest reference is 2011 GTC Asia Beijing, Thomas C. Schultess
- Fuhrer, O., C. Osuna, X. Lapillonne, T. Gysi, B. Cumming, M. Bianco, A. Arteaga, T. C. Schulthess,
  - 2014: Towards a performance portable, architecture agnostic implementation strategy for weather and climate models. 1 (1), 45-62 **doi: 10.14529/jsfi1401**
- Lapillonne, X., and O. Fuhrer, 2014:
  - Using compiler directives to port large scientific applications to GPUs: An example from atmospheric science. Parallel Processing Letters. 24, 1450003,
  - **doi: 10.1142/S0129626414500030**
- Fuhrer, O., T. Chadha, X. Lapillonne, D. Leutwyler, D. Luethi, C. Osuna, Ch. Schaer, T. Schulthess, H. Vogt, T. Hoefler, G. Kwasniewski, 2017
  - Near-global climate simulation at 1~km resolution on a GPU-accelerated supercomputer: establishing a performance baseline with COSMO 5.0. Geoscientific Model Development, 13 pages, in preparation.
- Leutwyler, D., D. Lüthi, N. Ban, O. Fuhrer, and C. Schär (2017),
  - Evaluation of the convection-resolving climate modeling approach on continental scales, J. Geophys. Res. Atmos., 122, 5237–5258, **doi:10.1002/2016JD026013**
- Leutwyler, D., O. Fuhrer, X. Lapillonne, D. Lüthi and C. Schär, 2016:
  - Towards European-Scale Convection-Resolving Climate Simulations. Geoscientific model development discussions, 2016, 1-34, doi: **10.5194/gmd-2016-119**

Who is using GPU for weather and climate?

UNIVERSITY OF LEEDS

- 2017 search of GMD papers for abstracts with GPU keyword
  - A search of their site reveals 9 papers since 2012
  - **ASAM, POP, POM, WRF(x2), GFDL-AM3, COSMO(x2) and NEMO**
  - The Princeton Ocean Model approach was to convert the whole code to CUDA-C
- NCAR & UCAR (CISL workshops in Sep each year)
  - MPAS, FV3, See Mark Govett (2016)
  - DWD Ulrich Schaettler (tinkering with radiation physics in COSMO)
  - MPAS (Korea Institute Science Technology)
  - UKMO UM and Lfric (PsyClone and GungHo)
- WRF developments
  - Michalakes, NREL as early as 2008 https://www.researchgate.net/publication/224316865
  - Mielikainen, Huang, 2012, “improved GPU/CUDA WRF + WSM5 cloud microphysics”
    - “IEEE Journal on Selected topics in applied earth observation and remote sensing vol.5 no.4”
  - Huang, 2015 doi:10.5194/gmd-8-2977-2015
    - GPU parallelization of planetary boundary layer (Yonsei scheme)
- Met office UM
  - Aware of some trials with several new architectures
  - Intel XeonPhi (Knight’s Landing)
  - NVIDIA GPU
  - Their focus has switched to the GW4 (Isembard) with ARM 64
    - Lower power requirements

- Introduction (MR 10 min)
- Case study (Sofya Titarenko 20 min)
  - Questions for Sofya
- Using GPUs on ARC3 (Mark Dixon 20 min)
  - Questions for Mark Dixon
- Round up (MR 5 min)

- Introduction (MR 10 min)
- Case study (Sofya Titarenko 20 min)
  - Questions for Sofya
- Using GPUs on ARC3 (Mark Dixon 20 min)
  - Questions for Mark Dixon
- Round up (MR 5 min)

Thank you Sofya and Mark

UNIVERSITY OF LEEDS

- Introduction (MR 10 min)
- Case study (Sofya Titarenko 20 min)
  - Questions for Sofya
- Using GPUs on ARC3 (Mark Dixon 20 min)
  - Questions for Mark Dixon
- Round up (MR 5 min)

Round-up

UNIVERSITY OF LEEDS

- I have given a brief high-level introduction to why we are talking about GPUs
  - Maturity
  - Fortran Double precision
  - Comparable costs to X86_64
  - Availability at Leeds
  - Potential codes already enabled
- Sofya presented us with a case study about wave propagation
  - Due to nature of her code almost 100% is passed to GPU
  - Some tips about targeting a GPU device
- Mark Dixon has shown us how we can access the GPUs at Leeds
  - User environment, batch job scripts, interrogate resource
- Main thrust in compilation of code is to choose one of
  - CUDA, OpenCL, OpenACC or CUDA Fortran
- Significant work needed to prepare code for GPU “acceleration”
  - Find someone who has done it already

Challenge of weather and climate codes.

- Flat profile
  - Often these codes have 100s of subroutines that do very similar amount of work
  - Only 3 or 4 that stand out and even then a maximum of 10% of the runtime.
- There is no single identifiable “kernel” so ideal speed up is only available in limited number of regions
- Target a subset of functionality instead
  - i.e. UKCA can be 30% of the runtime
  - I did this with the JWCRP project by identifying the whole aerosol subsystem and applying OpenMP around it
  - Early GPU systems would not have coped with the memory requirement.
  - Could work well for a modern GPU
    - (note earlier comment on Met Office direction).

Where can I get access to GPUs?

- UoLeeds ARC3
  - http://arc.leeds.ac.uk/systems/arc3/
  - Training at ARC on19th January http://arc.leeds.ac.uk/training/
- The Tier 2 system JADE http://www.jade.ac.uk/access/
  - "*All UK academic users are eligible to apply for time for Other applications through the EPSRC Tier 2 RAP (Resource Allocation Panel). Pump-priming time will also be available; please contact Prof Mike Giles in the first instance.*"
  - Primary partnership
    - Bristol, Edinburgh, Oxford and Southampton
  - Newer collaborations
    - Alan Turing Institute, Kings College London, Queen Mary London, Sheffield and UCL
  - Hosted at Hartree Daresbury (supplied by Atos)

Thank you to Sofya and Mark
Thank you for your attention!

Thank speakers

UNIVERSITY OF LEEDS

- Chair-person questions
  - Sofya
    - this effort you made with an old GPU
    - Will you be able to use a more modern GPU straight away
      - (will the CUDA part of the code need updating?)
  - MarkD
    - How busy are the queues? How long do I wait for a job to start?
    - Interactive more likely?
    - Limits of usage ?