Parallel Computing with GPUs

Dr Paul Richmond

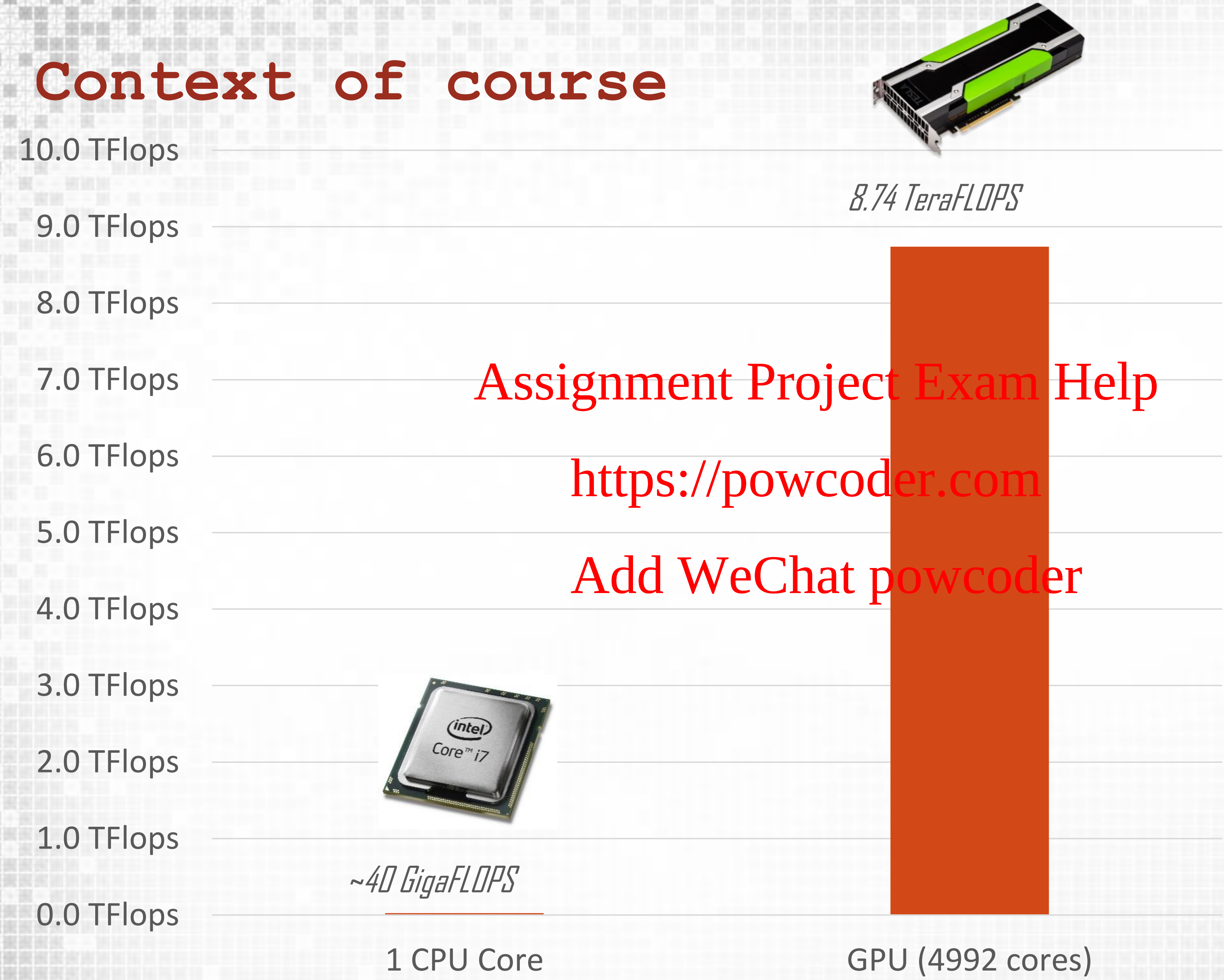
http://paulrichmond.shef.ac.uk/teaching/COM4521/

The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

- ❑ Context and Hardware Trends
- ❑ Supercomputing
- ❑ Software and Parallel Computing
- ❑ Course Outline

Context of course



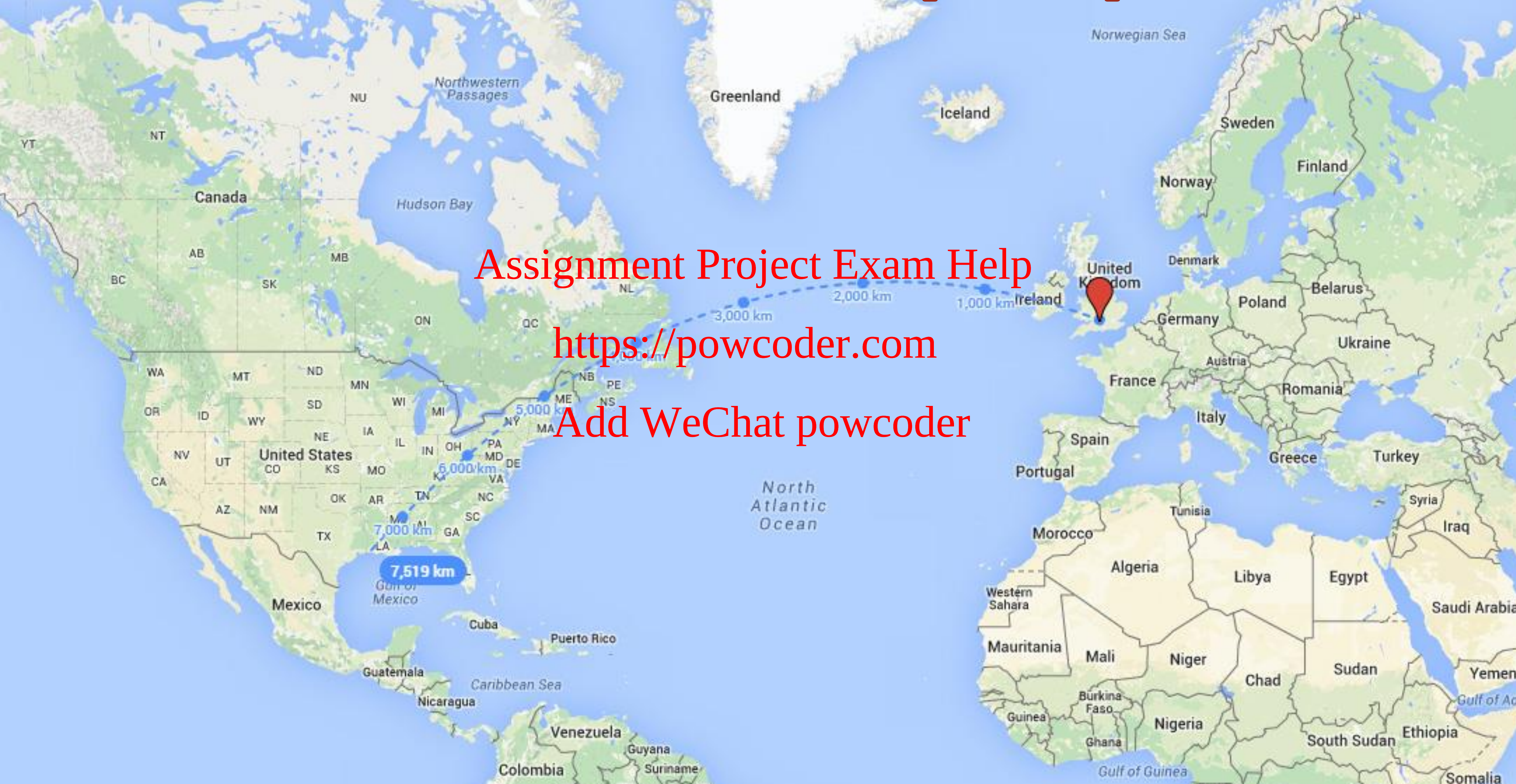
6 hours *CPU* time

vs.

1 minute *GPU* time

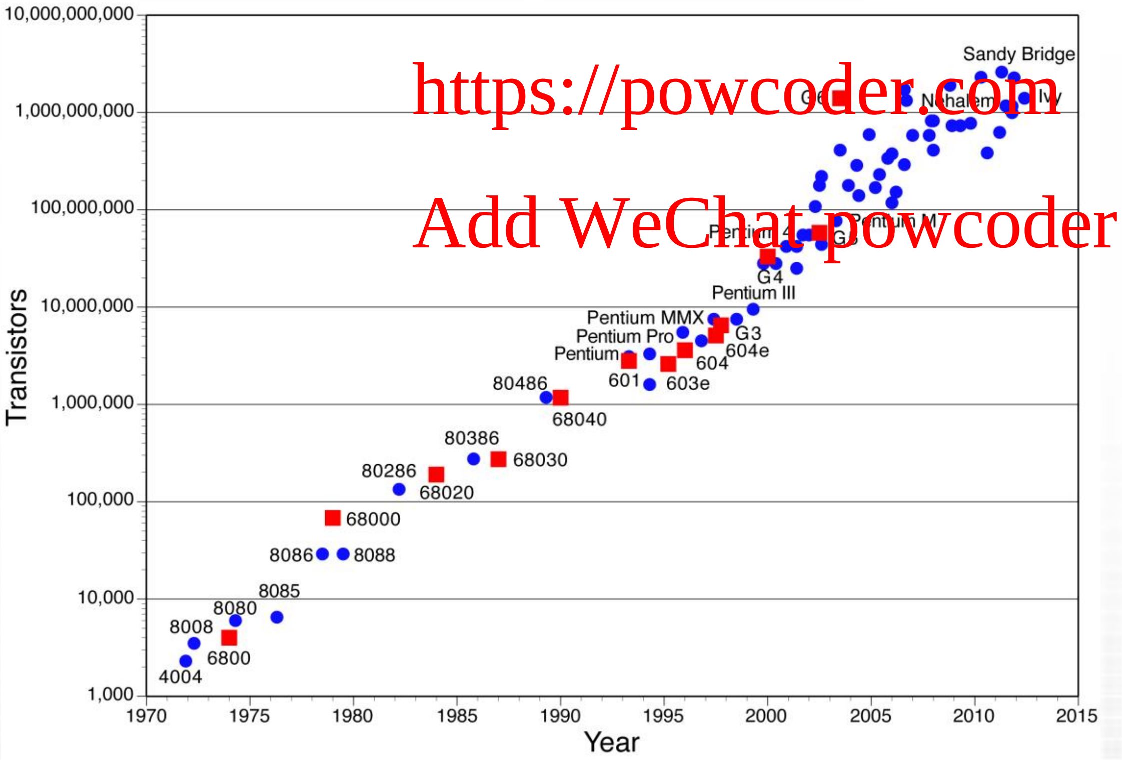
Scale of Performance

Scale of Performance: Titan Supercomputer



Transistors != performance

- Moores Law: A doubling of transistors every couple of years
  - Not a law actually an observation
  - Doesn’t actually say anything about performance



Dennard Scaling

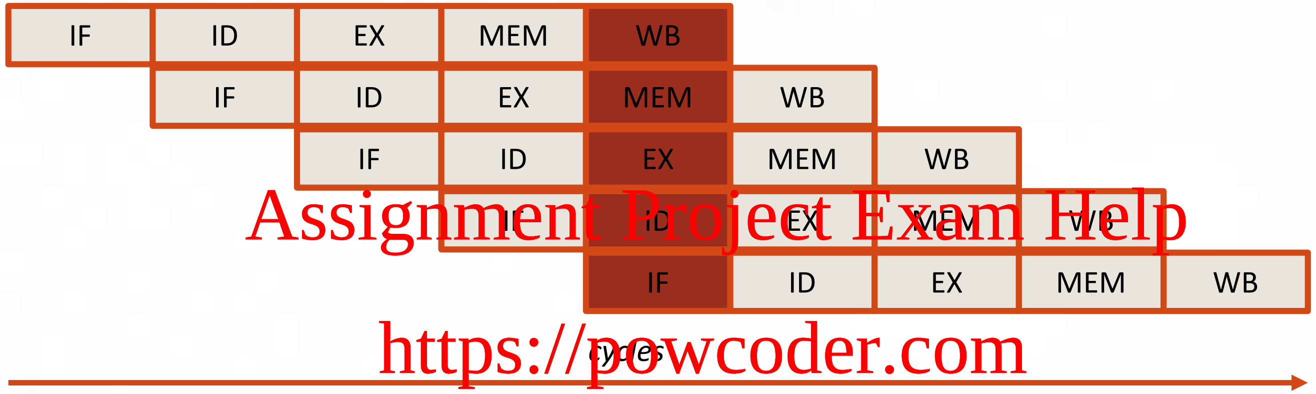
"As transistors get smaller their power density stays constant"

Power = Frequency x Voltage^2

- Performance improvements for CPUs traditionally realised by increasing frequency

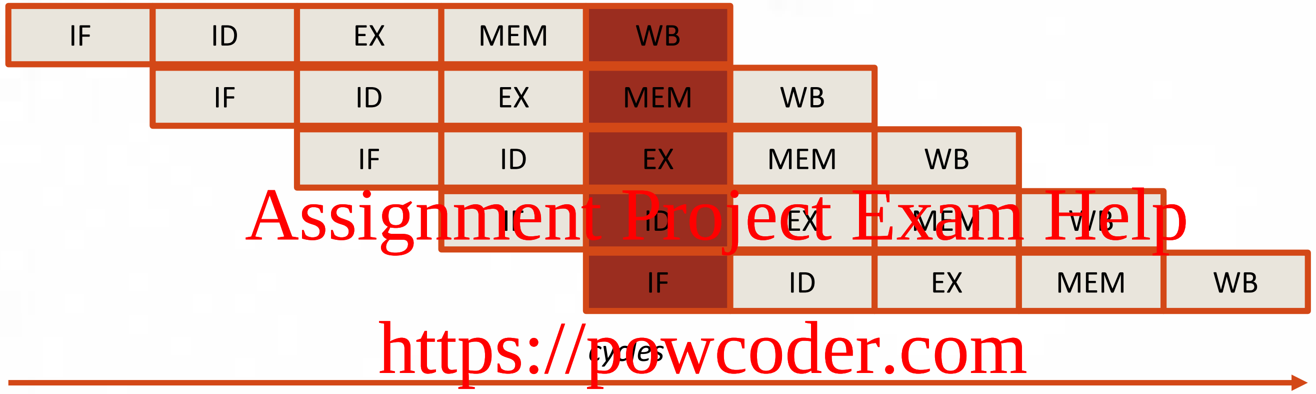
- Decrease voltage to maintain a steady power
  - Only works so far
- Increase Power
  - Disastrous implications for cooling

Instruction Level Parallelism



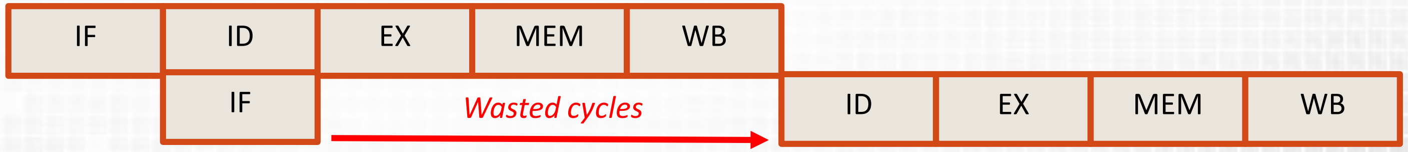
- ❑Transistors used to build more complex architectures
- ❑Use pipelining to overlap instruction execution

Instruction Level Parallelism

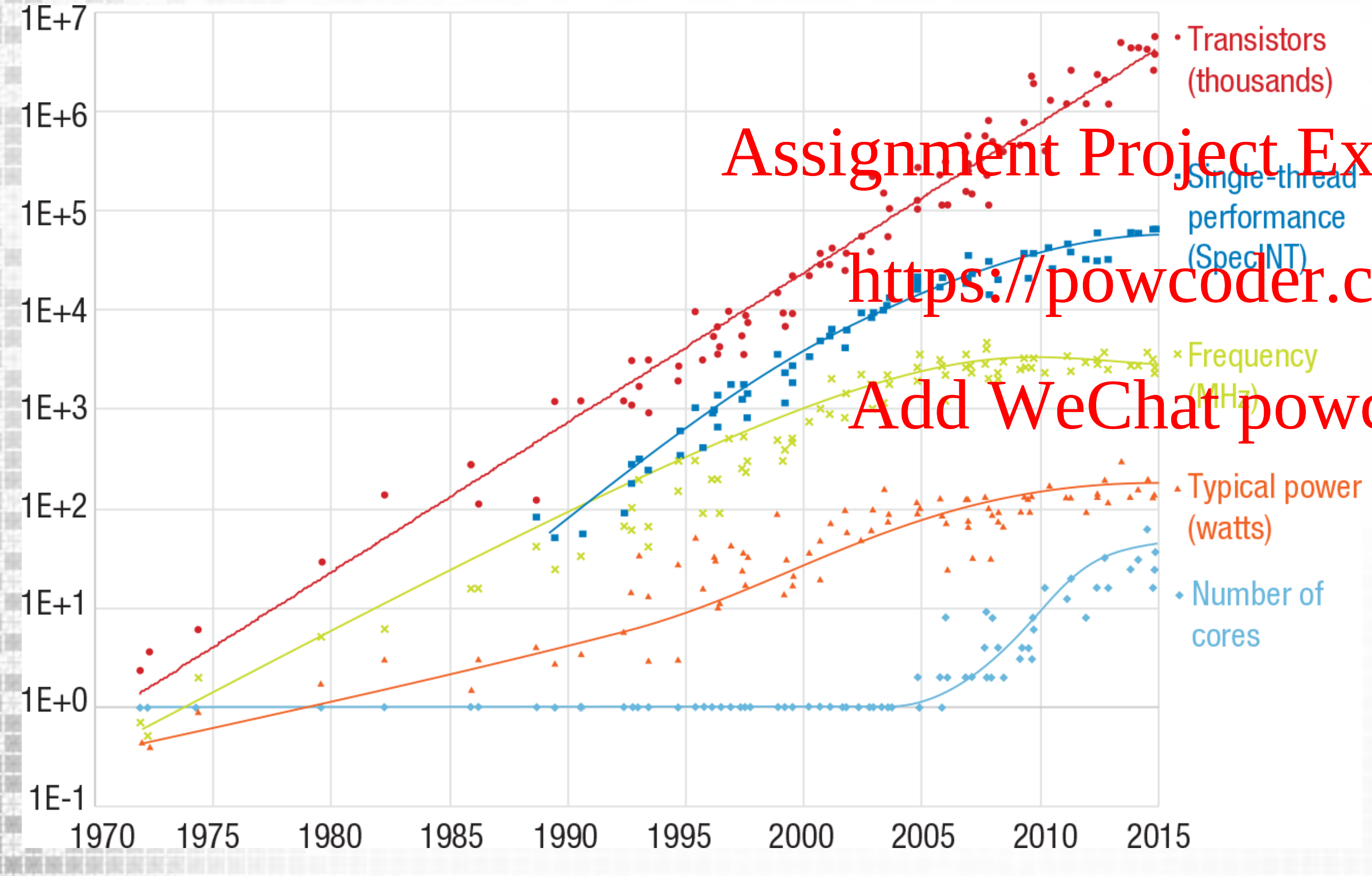


- Transistors used to build more complex architectures
- Use pipelining to overlap instruction execution

```
add 1 to R1
copy R1 to R2
```



Golden Era of Performance



Adapting to Thrive in a New Economy of Memory Abundance, K Bresniker et al.

- ❑90s saw great improvements to single CPU performance
  - ❑1980s to 2002: 100% performance increase every 2 years
  - ❑2002 to now: ~40% every 2 years

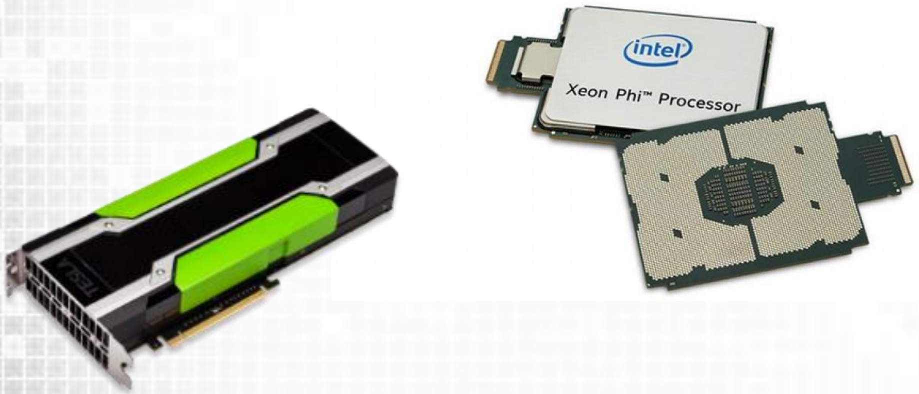
GPU RESEARCH CENTER

Why More Cores?

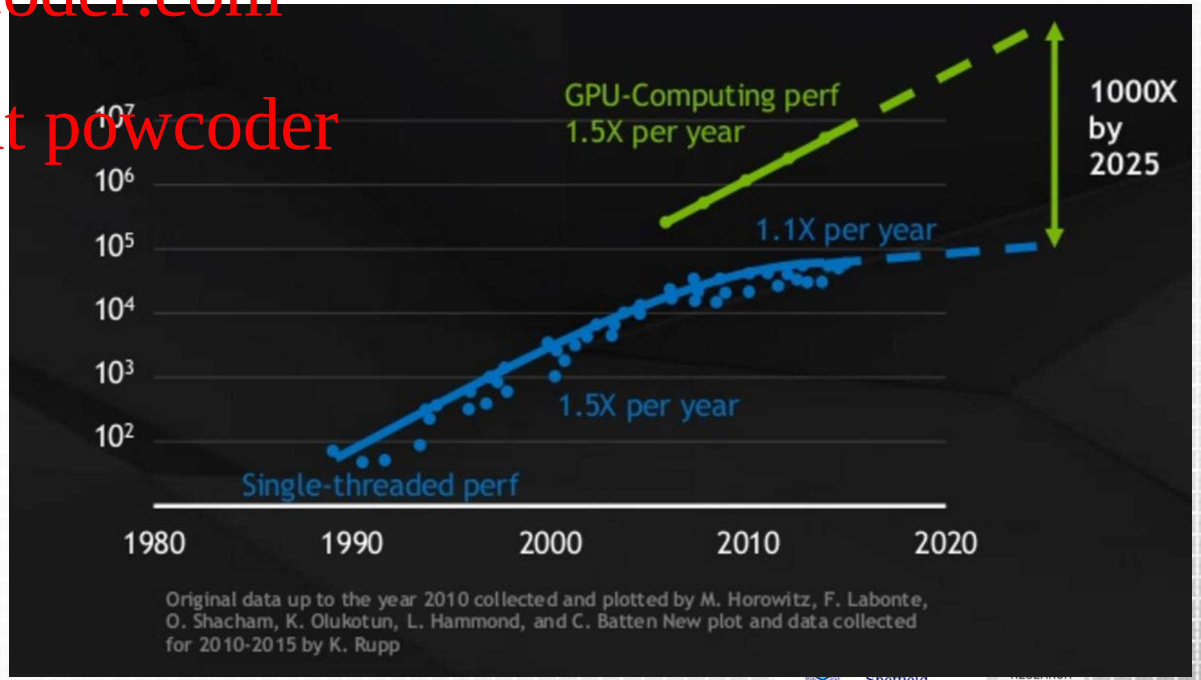
- ❑Use extra transistors for multi/many core parallelism
  - ❑More operations per clock cycle
  - ❑Power can be kept low
  - ❑Processor designs can be simple – shorter pipelines (RISC)

GPUs and Many Core Designs

- ❑Take the idea of multiple cores to the extreme (many cores)
- ❑Dedicate more die space to compute
  - ❑At the expense of branch prediction, out of order execution, etc.
- ❑Simple, Lower Power and Highly Parallel
  - ❑Very effective for HPC applications

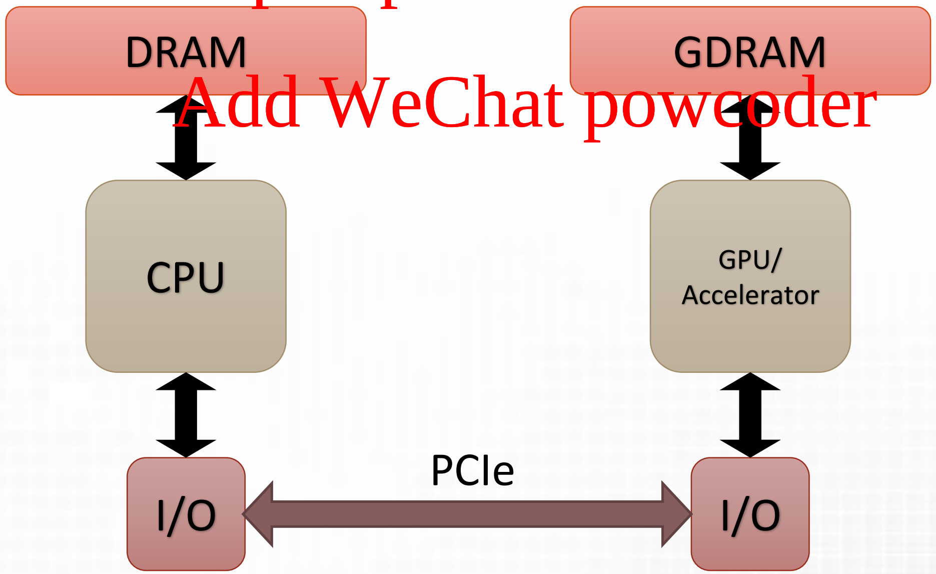


From GTC 2017 Keynote Talk, NVIDIA CEO Jensen Huang



Accelerators

- Problem: Still require OS, IO and scheduling
- Solution: "Hybrid System",
  - CPU provides management and
  - "Accelerators" (or co-processors) such as GPUs provide compute power



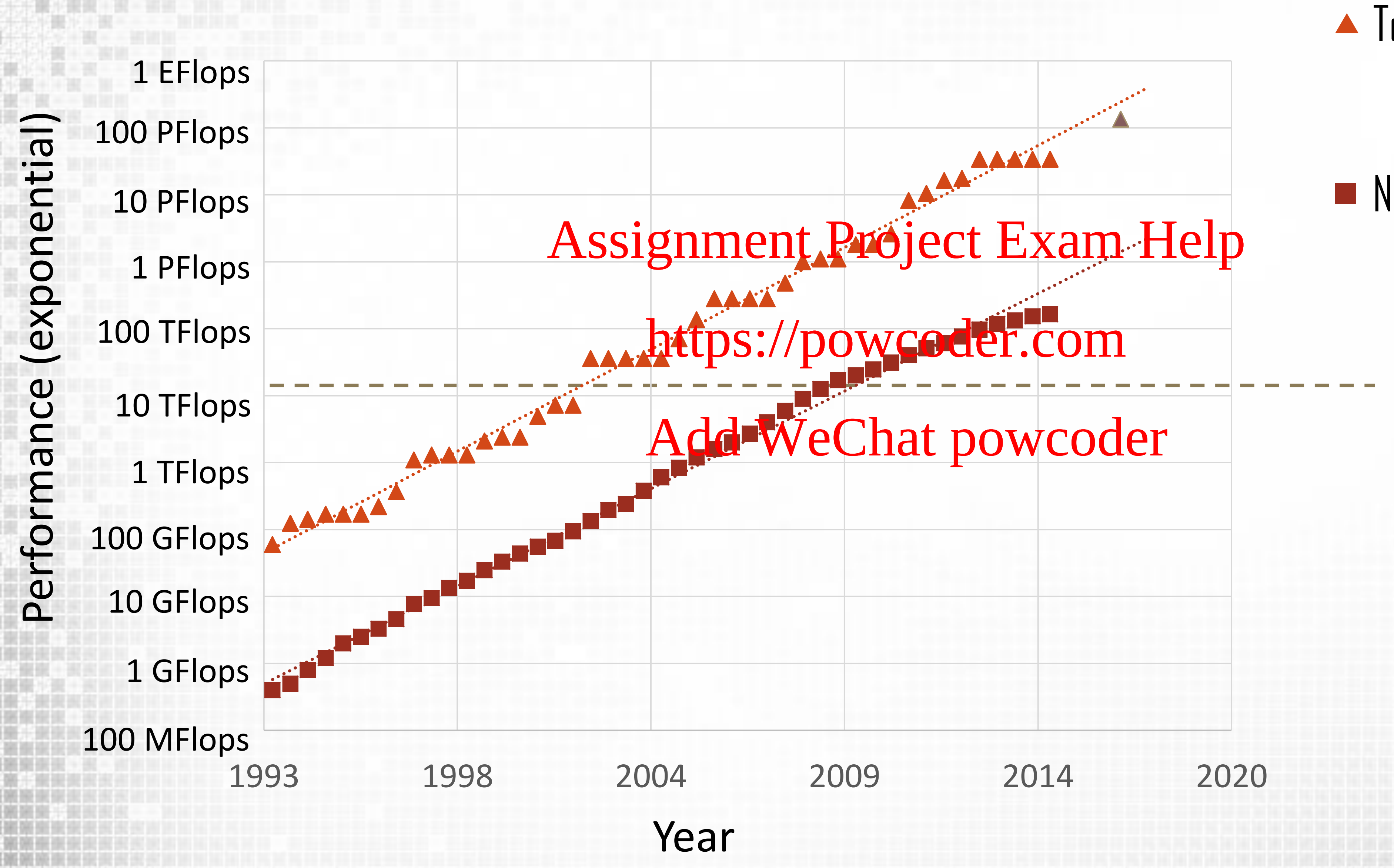
Types of Accelerator

- GPUs
  - Emerged from 3D graphics but now specialised for HPC
  - Readily available in workstations

- Xeon Phis
  - Many Integrated Cores (MIC) architecture
  - Based on Pentium 4 design (x86) with wide vector units
  - Closer to traditional multicore
  - Simpler programming and compilation

- ❑Context and Hardware Trends
- ❑Supercomputing
- ❑Software and Parallel Computing
- ❑Course Outline

Top Supercomputers



Volta V100 (15TFLOPS SP)

Supercomputing Observations

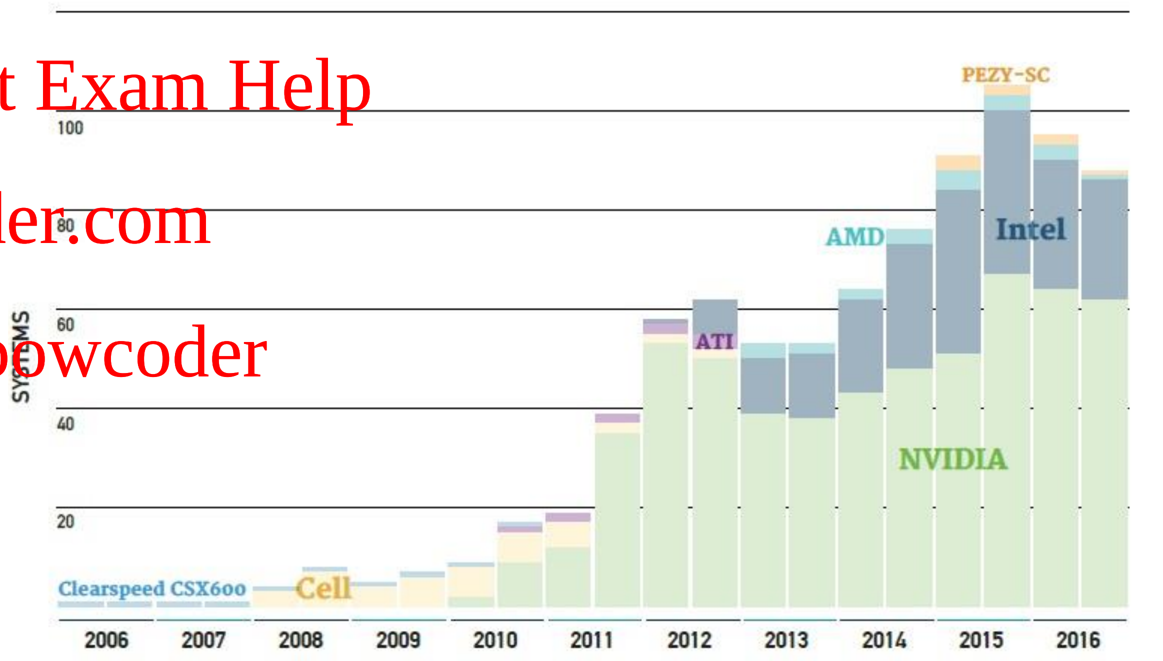
- Exascale computing
  - 1 Exaflop = 1M Gigaflops
  - Estimated for 2020
- Pace of change
  - Desktop GPU top supercomputer in 2002
  - A desktop with a GPU would be in Top 500 in 2008
  - A Teraflop of performance took 1MW in 2000
- Extrapolating the trend
  - Current gen top500 on every desktop in < 10 years

Trends of HPC

- Improvements at individual computer node level are greatest
  - Better parallelism
  - Hybrid processing
  - 3D fabrication
- Communication costs are increasing
  - Memory per core is reducing

Supercomputing Observations

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Supercomputing Center in Wuxi China	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
3	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
4	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939
6	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	13,554.6	24,913.5	2,719
7	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
8	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 Cray Inc.	206,720	9,779.0	15,988.0	1,312



https://www.nextplatform.com/2016/11/14/closer-look-2016-top-500-supercomputer-rankings/

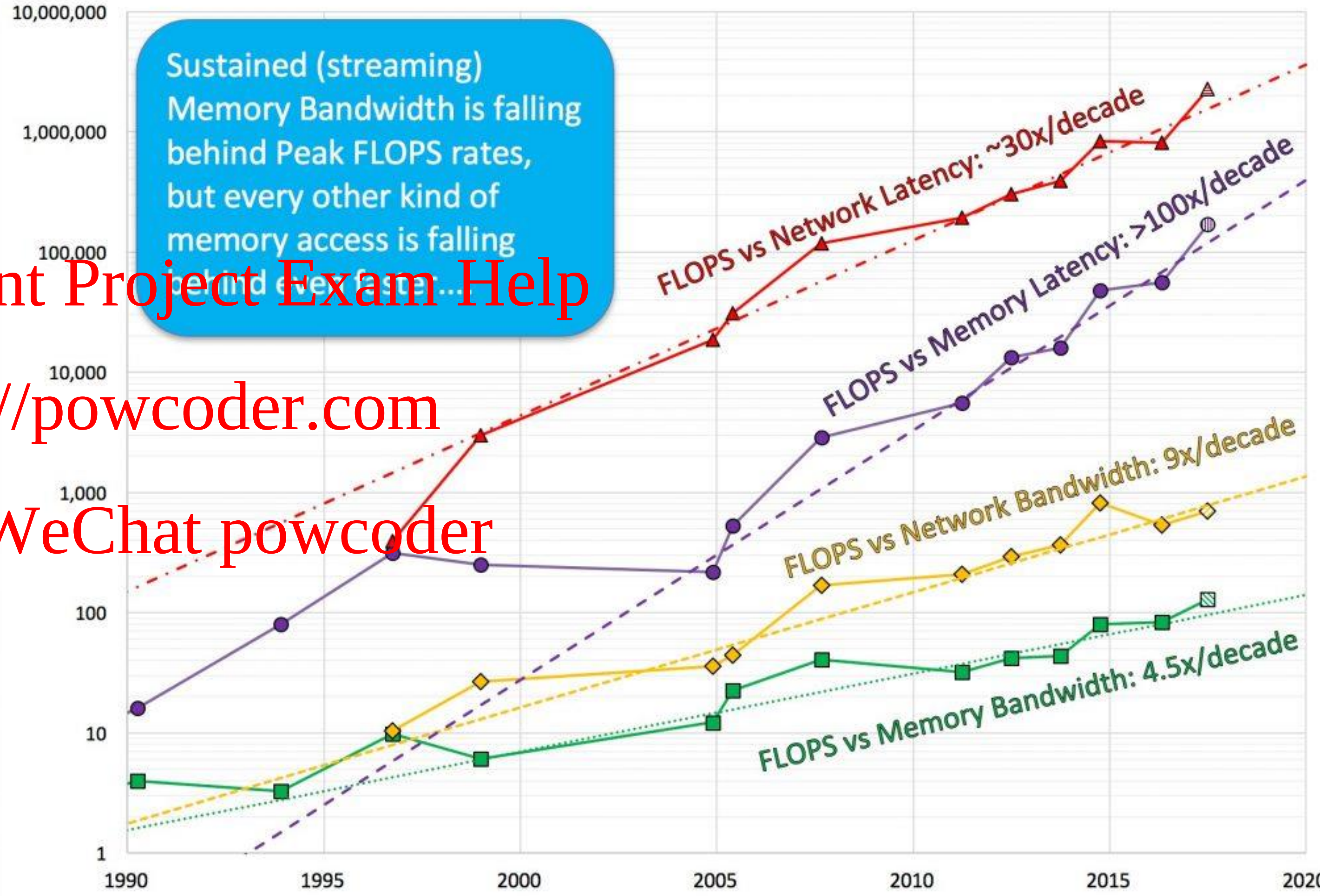
Green 500

- Top energy efficient supercomputers

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband FDR, NVIDIA Tesla P100 , ExaScalar Yahoo Japan Corporation Japan	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC National Institute of Advanced Industrial Science and Technology Japan	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Fujitsu Center for Advanced Intelligence Project, RIKEN Japan	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell University of Cambridge United Kingdom	21,240	1,193.0	114	10.428
6	3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	2,272	10.398
7	69	**Gyoukou** - ZettaScaler-2.0 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , ExaScalar Japan Agency for Marine -Earth Science and Technology Japan	3,176,000	1,677.1	164	10.226
8	220	**Research Computation Facility for GOSAT-2 (RCF2)** - SGI	16,320	770.4	79	9.797

HPC Observations

- ❑Improvements at individual computer node level are greatest
  - ❑Better parallelism
  - ❑Hybrid processing
  - ❑3D fabrication
- ❑Communication costs are increasing
  - ❑Memory per core is reducing
- ❑Throughput > Latency



http://sc16.supercomputing.org/2016/10/07/sc16-invited-talk-spotlight-dr-john-d-mccalpin-presents-memory-bandwidth-system-balance-hpc-systems/

The University Of Sheffield. GPU RESEARCH CENTER

- ❑ Context and Hardware Trends
- ❑ Supercomputing
- ❑ Software and Parallel Computing
- ❑ Course Outline

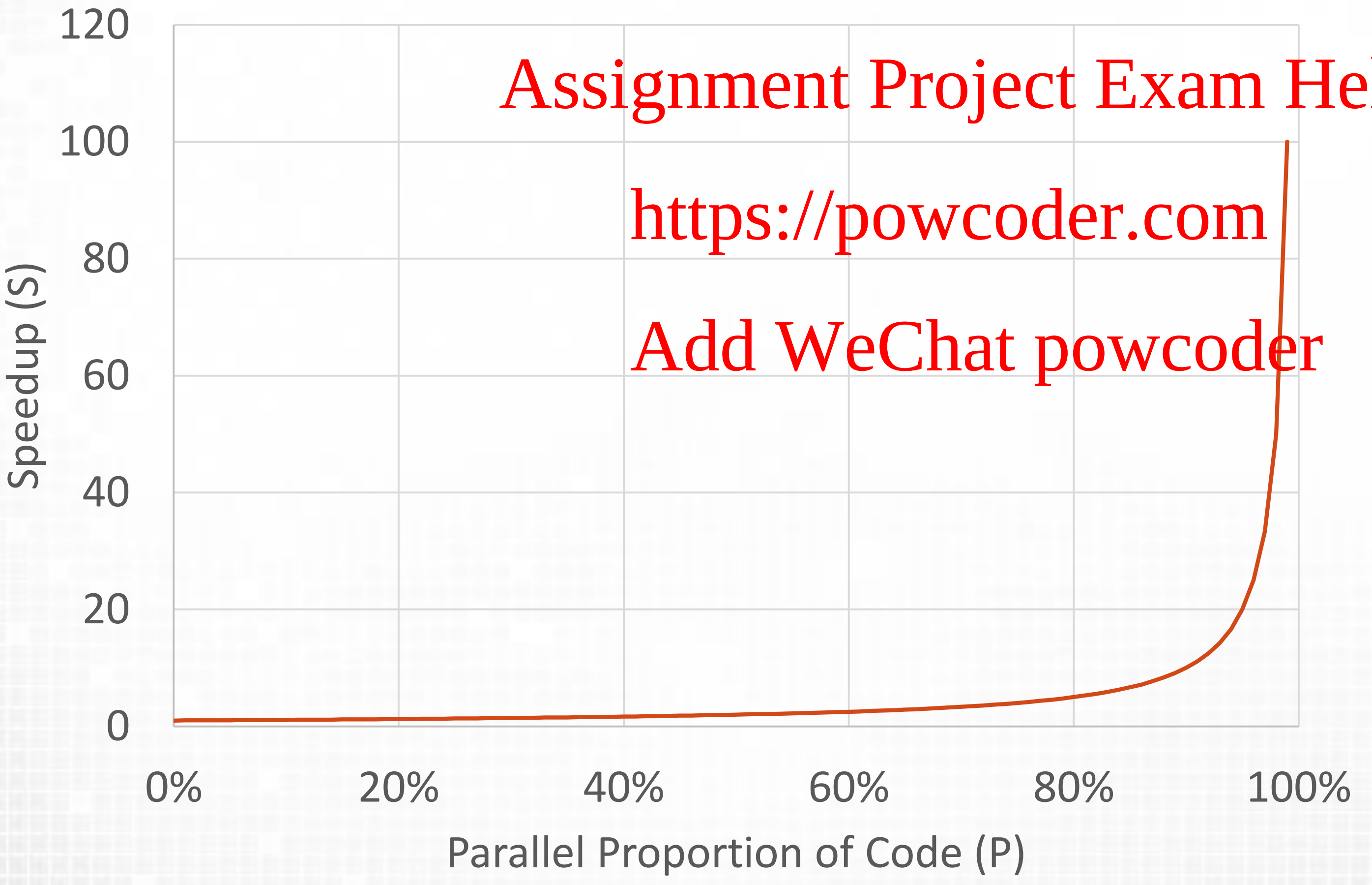
Software Challenge

- ❑ How to use this hardware efficiently?

- ❑ Software approaches
  - ❑ Parallel languages: some limited impact but not as flexible as sequential programming
  - ❑ Automatic parallelisation of serial code: >30 years of research hasn't solved this yet
  - ❑ **Design software with parallelisation in mind**

Amdahl's Law

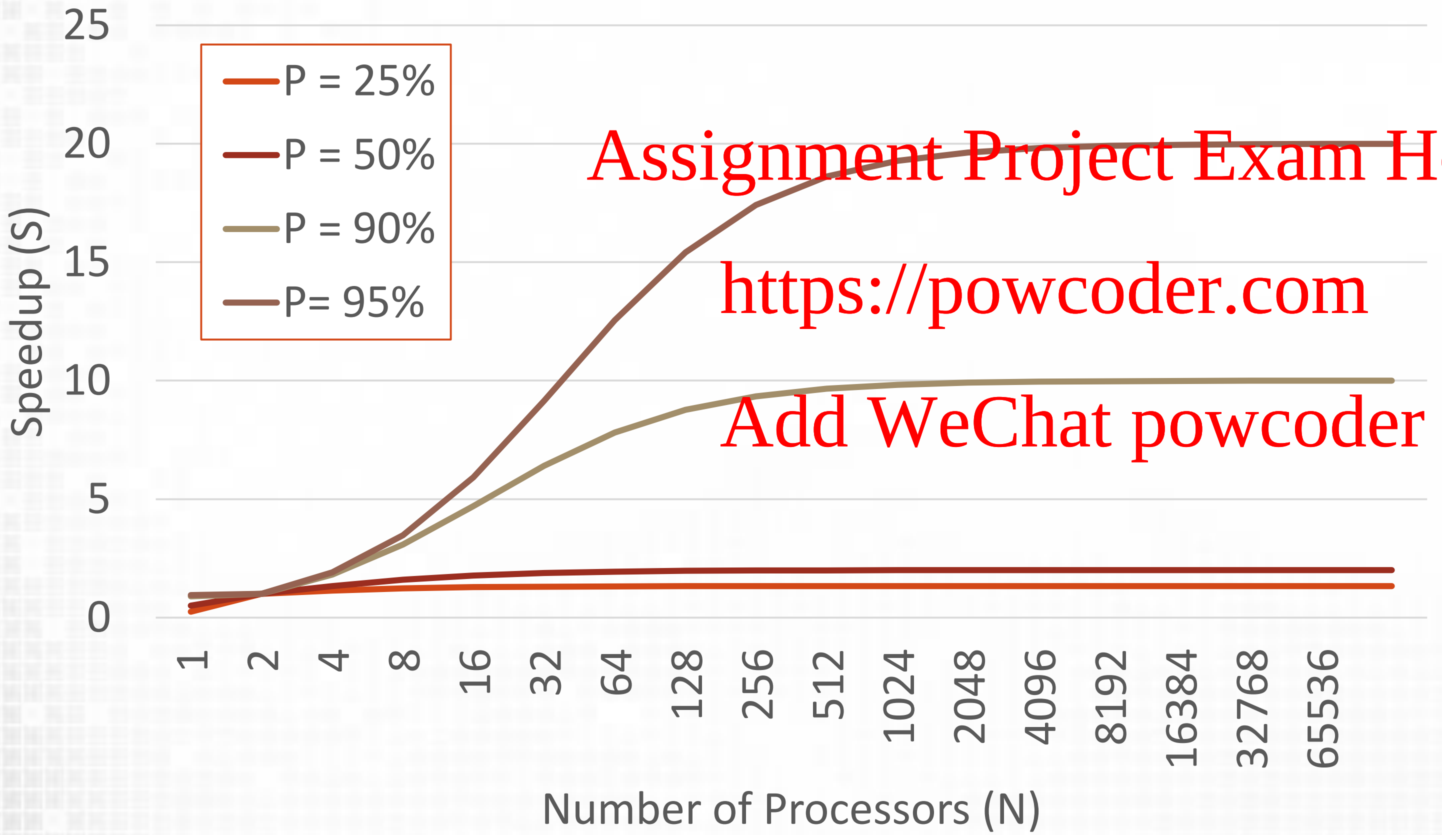
- Speedup of a program is limited by the proportion than can be parallelised



$$Speedup\ (S) = \frac{1}{1 - P}$$

Amdahl's Law cont.

- ❑Addition of processing cores gives diminishing returns



$$Speedup\ (S) = \frac{1}{\frac{P}{N} - (1 - P)}$$

Parallel Programming Models

- ❑Distributed Memory
  - ❑Geographically distributed processors (clusters)
  - ❑Information exchanged via messages
- ❑Shared Memory
  - ❑Independent tasks share memory space
  - ❑Asynchronous memory access
  - ❑Serialisation and synchronisation to ensure correctness
  - ❑No clear ownership of data
  - ❑Not necessarily performance oriented

Types of Parallelism

- ❑ Bit-level
  - ❑ Parallelism over size of word, 8, 16, 32, or 64 bit.
- ❑ Instruction Level (ILP)
  - ❑ Pipelining

- ❑ Task Parallel

  - ❑ Program consists of many independent tasks
  - ❑ Tasks execute on asynchronous cores

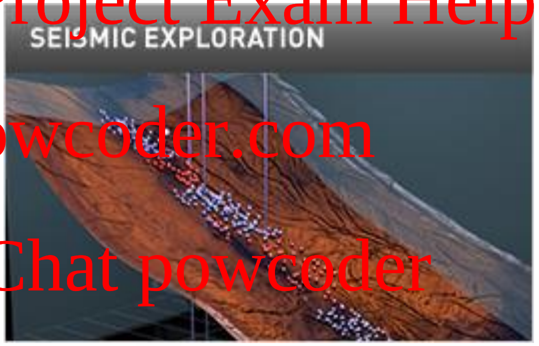
- ❑ Data Parallel
  - ❑ Program has many similar threads of execution
  - ❑ Each thread performs the same behaviour on different data

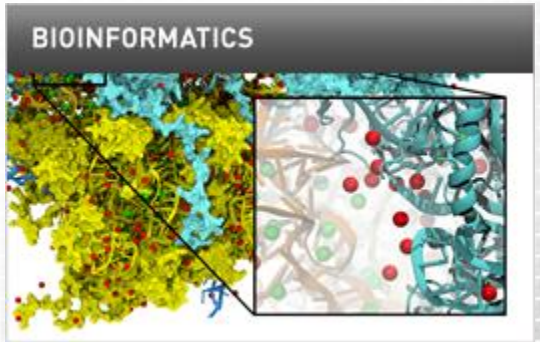
The University Of Sheffield.

NVIDIA GPU RESEARCH CENTER

Implications of Parallel Computing

- ❑Performance improvements
  - ❑Speed
  - ❑Capability (i.e. scale)





- ❑Context and Hardware Trends
- ❑Supercomputing
- ❑Software and Parallel Computing
- ❑Course Outline

COM4521/6521 specifics

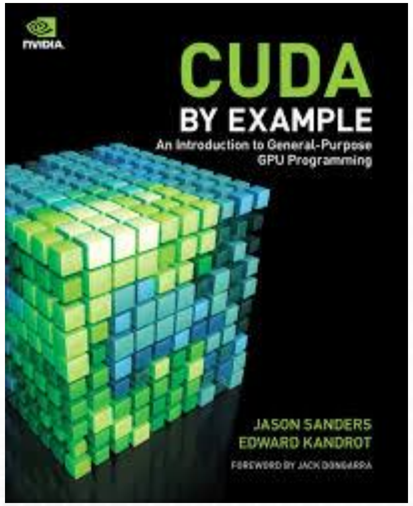
- Designed to give insight into parallel computing
  - Specifically with GPU accelerators
  - Knowledge transfers to all many core architectures
- What you will learn
  - How to program in C and manage memory manually
  - How to use OpenMP to write programs for multi-core CPUs
  - What a GPU is and how to program it with the CUDA language
  - How to think about problems in a highly parallel way
  - How to identify performance limitations in code and address them

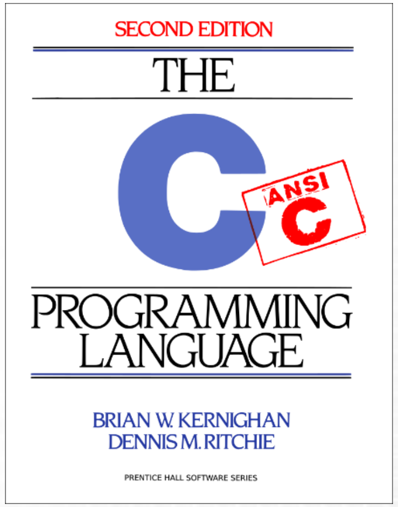
Course Mailing List

- A google group for the course has been set up
  - You have already been added if you were registered 01/02/2018
- Mailing list uses;
  - Request help outside of lab classes
  - Find out if a lecture has changed
  - Want to participate in discussion on course content
- https://groups.google.com/a/sheffield.ac.uk/forum/#!forum/com4521-group

Learning Resources

- Course website: http://paulrichmond.shef.ac.uk/teaching/COM4521/
- Recommended Reading:
  - Edward Kandrot, Jason Sanders, "CUDA by Example: An Introduction to General-Purpose GPU Programming", Addison Wesley 2010.
  - Brian Kernighan, Dennis Ritchie, “The C Programming Language (2nd Edition)”, Prentice Hall 1988.





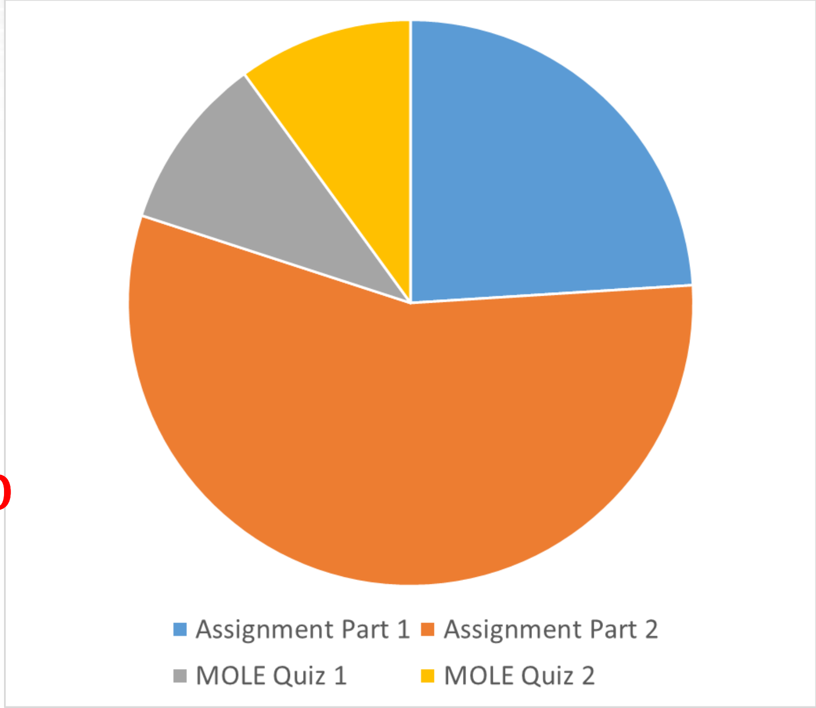
Timetable

- 2 x 1 hour lecture per week (back to back)
  - Monday 15:00 until 17:00 Broad Lane Lecture Theater 11
  - Week 5 first half of the lecture will be in DIA-LT09 (Lecture Theatre 9)
  - Week 5 second half of the lecture will be MOLE quiz in DIA-206 (Compute room 4)
- 1 x 2 hour lab per week
  - Tuesday 9:00 until 11:00 Diamond DIA-206 (Compute room 4)
  - Week 10 first half of the lab will be an assessed MOLE quiz DIA-206 (Compute room 4)

- Assignment
  - Released in two parts
  - Part 1
    - Released week 3
    - Due for hand in on Tuesday week 7 (20/03/2018) at 17:00
    - Feedback after Easter.
  - Part 2
    - Released week 6
    - Due for hand in on Tuesday week 12 (15/05/2018) at 17:00

Course Assessment

- 2 x Multiple Choice quizzes on MOLE (10% each)
  - Weeks 5 and 10
- An assignment (80%)
  - Part 1 is 30% of the assignment total
  - Part 2 is 70% of the assignment total
- For each assignment part
  - Half of the marks are for the program and half for a written report
  - Will require understanding of why you have implemented a particular technique
  - Will require benchmarking, profiling and explanation to demonstrate that you understand the implications of what you have done



The University Of Sheffield.
NVIDIA GPU RESEARCH CENTER

Lab Classes

- 2 hours every week
  - Essential in understanding the course content!
  - Do not expect to complete all exercises within the 2 hours
- Coding help from lab demonstrators Robert Chisholm and John Charlton:
  - http://staffwww.dcs.shef.ac.uk/people/R.Chisholm/
  - http://www.dcs.shef.ac.uk/cgi-bin/makeperson?J.Charlton
- Assignment and lab class help questions should be directed to the google discussion group


Feedback

- After each teaching week you MUST submit the lab register/feedback form
  - This records your engagement in the course
  - Ensures that I can see what you have understood and not understood
  - Allows us to revisit any concepts ideas with further examples
  - This only works if you are honest!
- Submit this once you have finished with the lab exercises
- Your feedback will be used to clarify topics which are assessed in the assignments
- Lab Register Link: https://goo.gl/0r73gD
- Additional feedback from assignment and MOLE quizzes

Machines Available

- Diamond Compute Labs
  - Visual Studio 2017
  - NVIDIA CUDA 9.1
- VAR Lab
  - CUDA enabled machines – same spec as Diamond high spec compute room
- ShARC
  - University of Sheffield HPC system
  - You will need an account (see HPC docs website)
  - Select number of GPU nodes available (see gpucomputing.shef.ac.uk)
  - Special short job queue will be made availble
- Your own machine
  - Must have a NVIDIA GPU for CUDA exercises
  - Virtual machines not an option
  - **IMPORTANT**: Follow the websites guidance for installing Visual Studio

Summary

- Parallelism is already here in a big way
  - From mobile to workstation to supercomputers
- Parallelism in hardware
  - It's the only way to use increasing number of transistors
  - Trend is for increasing parallelism
- Supercomputers
  - Increased dependency on accelerators
  - Accelerators are greener
- Software approaches
  - Shared and distributed memory models differ
  - Programs must be highly parallel to avoid diminishing returns