HPC ARCHITECTURES

Darren White

d.white@epcc.ed.ac.uk

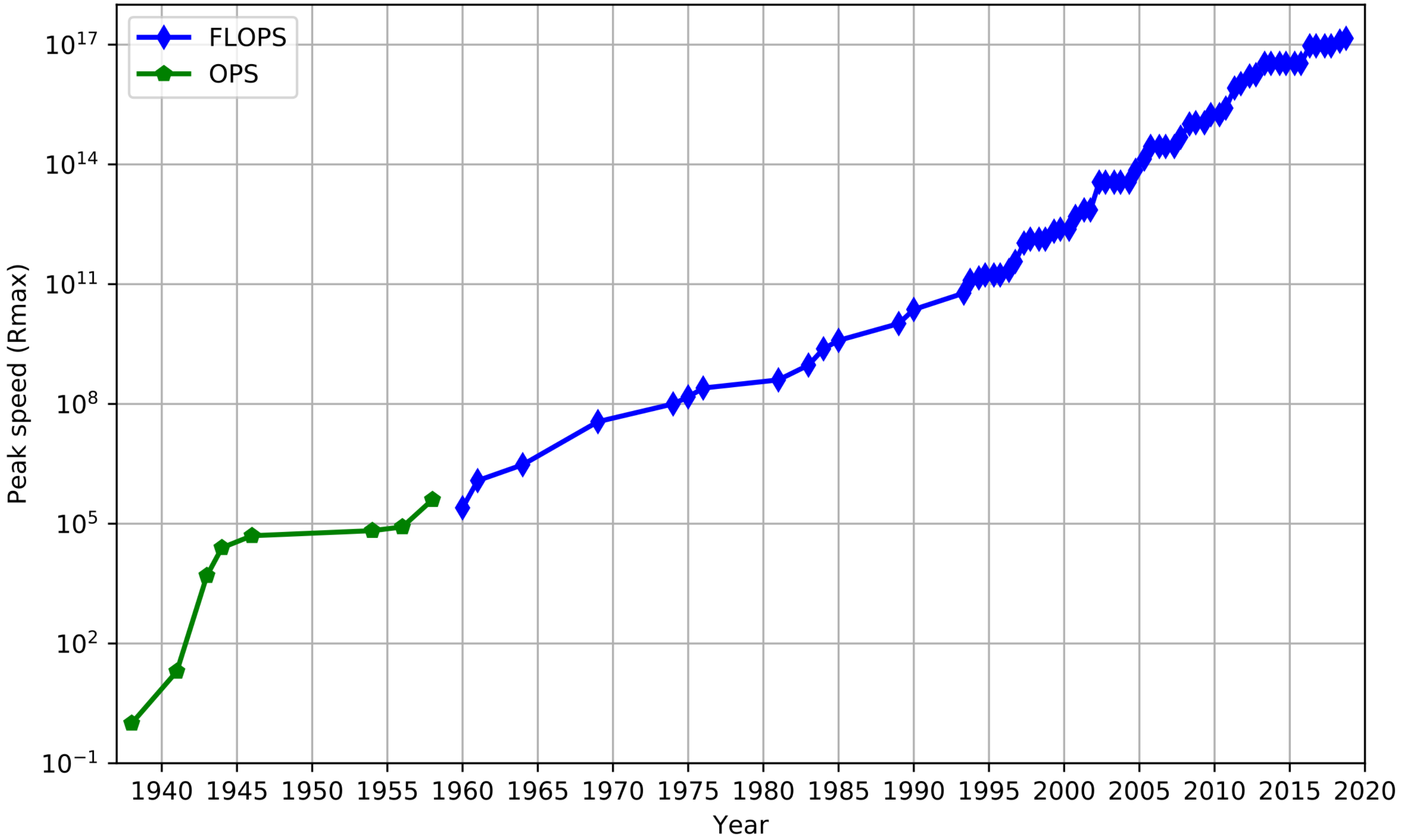
Overview

- HPC Architecture course aims to cover:
  - Basic components of HPC systems: processors, memory, interconnect, storage.
  - Classification of architectures: SIMD/MIMD, shared vs distributed memory, clusters
  - System software: OSs, processes, threads, scheduling.
  - Brief history of HPC systems, including Moore's Law.
  - CPU design: functional units, instructions sets, pipelining, branch prediction, ILP (superscalar, VLIW, SIMD instructions), multithreading.
  - Caches: operation and design features
  - Memory: operation and design features, including cache coherency and consistency
  - Multicore CPUs, including cache and memory hierarchy
  - GPGPUs: operation and design features
  - Interconnects: operation and design features
  - Filesystems and associated hardware
  - Data focussed hardware and system integration
  - Batch systems
  - Energy and power performance and monitoring
  - Current HPC architectures

Overview

- Timetable:
  - Monday 11:10-12:00 – lecture (Bayes G.03)
  - Tuesday 14:10-15:00 – practical (Appleton 5.05)
  - Friday 11:10-12:00 – lecture (Bayes G.03)
- Assessed through exam
  - 4 question, 25 marks each question
- Reading resources
  - **Introduction to High Performance Computing for Scientists and Engineers –** Georg Hager, Gerhard Wellein
  - **Introduction to High Performance Scientific Computing -** Victor Eijkhout, http://www.tacc.utexas.edu/~eijkhout/Articles/EijkhoutIntroToHPC.pdf

Performance Trend



This graph is borrowed from Wikipedia ©Lucas wilkins

FLOPS

- Yotta: 10^{24}
- Zetta: 10^{21}
- Exa: 10^{18}
- Peta: 10^{15}
- Tera: 10^{12}
- Giga: 10^{9}
- Mega: 10^{6}
- Kilo: 10^{3}

Quantifying Performance

- Serial computing concerned with complexity
  - how execution time varies with problem size *N*
  - adding two arrays (or *vectors*) is *O(N)*
  - matrix times vector is $O(N^2)$, matrix-matrix is $O(N^3)$
- Look for clever algorithms
  - naïve sort is $O(N^2)$
  - divide-and-conquer approaches are *O(N log (N))*
- Parallel computing *also* concerned with scaling
  - how time varies with number of processors *P*
  - different algorithms can have different scaling behaviour
  - but always remember that we are interested in minimum time!

Performance Measures

- *T(N,P)* is execution time for size *N* on *P* processors

- Speedup
  - typically *S(N,P)* < *P*

$$S(N,P) = \frac{T(N,1)}{T(N,P)}$$

- Parallel Efficiency
  - typically *E(N,P)* < 1

$$E(N,P) = \frac{S(N,P)}{P} = \frac{T(N,1)}{PT(N,P)}$$

- Serial Efficiency
  - typically E(N) <= 1

$$E(N) = \frac{T_{best}(N)}{T(N,1)}$$

Parallel Scaling

Scaling describes how the runtime of a parallel application changes as the number of processors is increased

Can investigate two types of scaling:

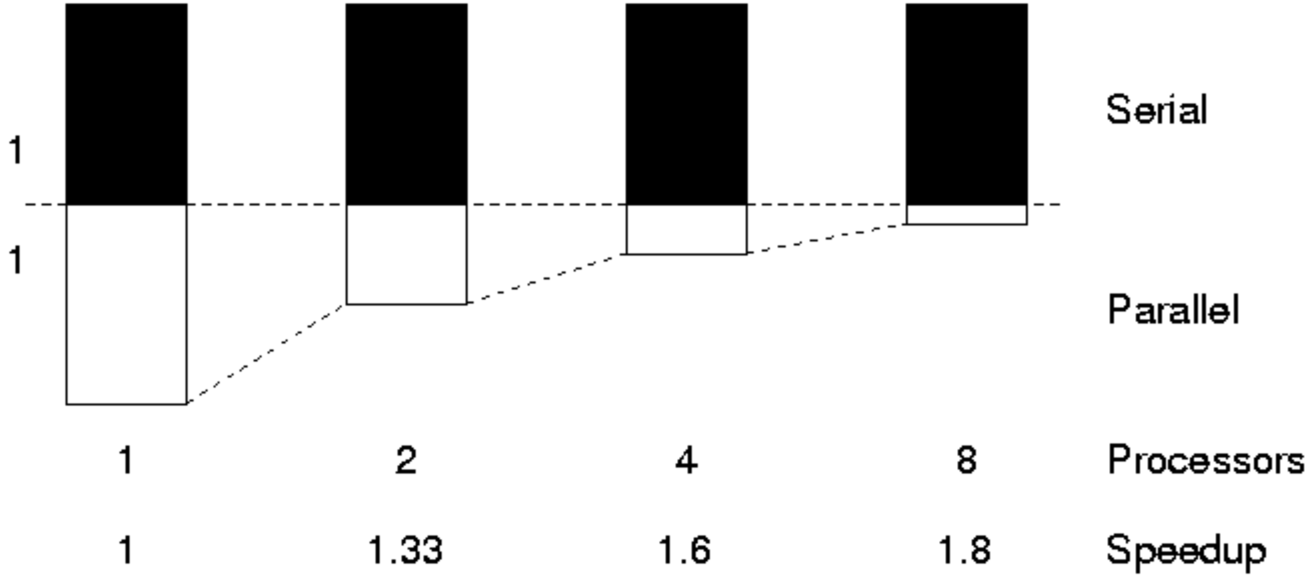
- Strong Scaling (increasing P, constant N):
  - problem size/complexity stays the same as the number of processors increases, decreasing the work per processor
- Weak Scaling (increasing P , increasing N):
  - problem size/complexity increases at the same rate as the number of processors, keeping the work per processor the same

The Serial Component

- Amdahl’s law

> *“the performance improvement to be gained by parallelisation is limited by the proportion of the code which is serial”*

Gene Amdahl, 1967



Amdahl’s law

- Assume a fraction α is completely serial
  - time is sum of serial and potentially parallel

- Parallel time

$$T(N,P) = \alpha T(N,1) + \frac{(1-\alpha)T(N,1)}{P}$$

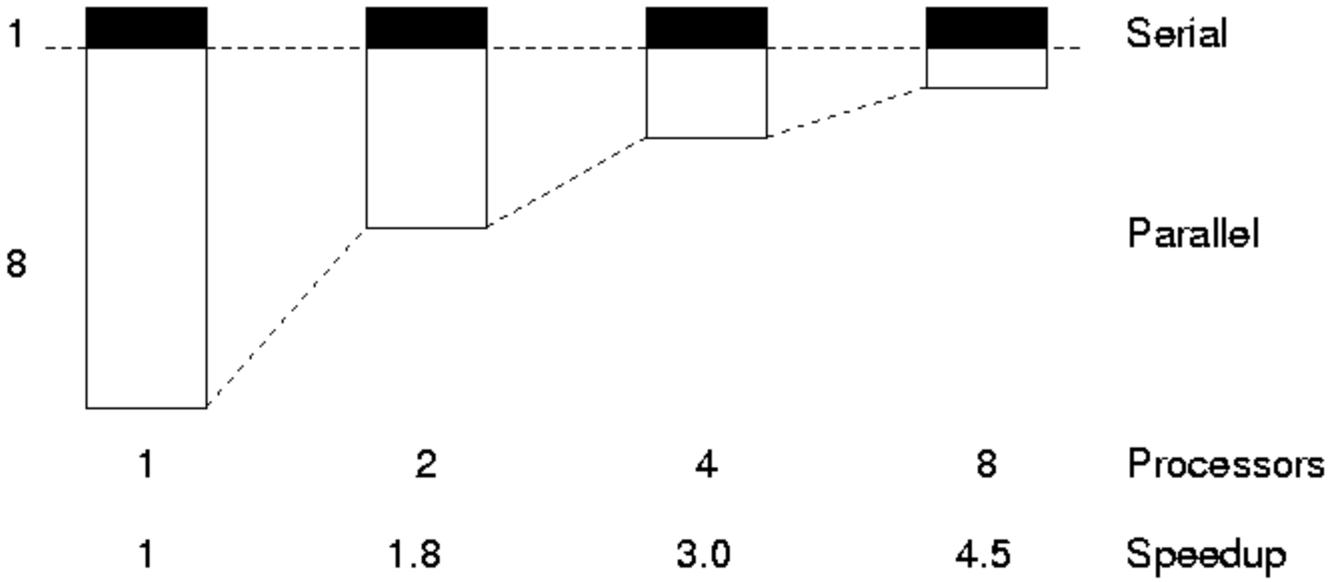
- Parallel speedup

$$S(N,P) = \frac{T(N,1)}{T(N,P)} = \frac{P}{\alpha P + (1-\alpha)}$$

  - for $\alpha = 0$, $S = P$ as expected (i.e. E = 100%)
  - otherwise, speedup limited by $1/\alpha$ for any P
  - impossible to effectively utilise large parallel machines?

Gustafson’s Law

- Need larger problems for larger numbers of CPUs



Utilising Large Parallel Machines

- Assume parallel part is *O(N)*, serial part is *O(1)*
  - time $$T(N,P) = T_{serial}(N,P) + T_{parallel}(N,P) \\ = \alpha T(1,1) + \frac{(1-\alpha)NT(1,1)}{P}$$
  - speedup $$S(N,P) = \frac{T(N,1)}{T(N,P)} = \frac{\alpha + (1-\alpha)N}{\alpha + (1-\alpha)\frac{N}{P}}$$
- Scale problem size with CPUs, ie set *N* = *P* ***Weak Scaling***
  - speedup $$S(P,P) = \alpha + (1-\alpha)P$$
  - efficiency $$E(P,P) = \frac{\alpha}{P} + (1-\alpha)$$
- Maintain constant efficiency (1-α) for large *P*

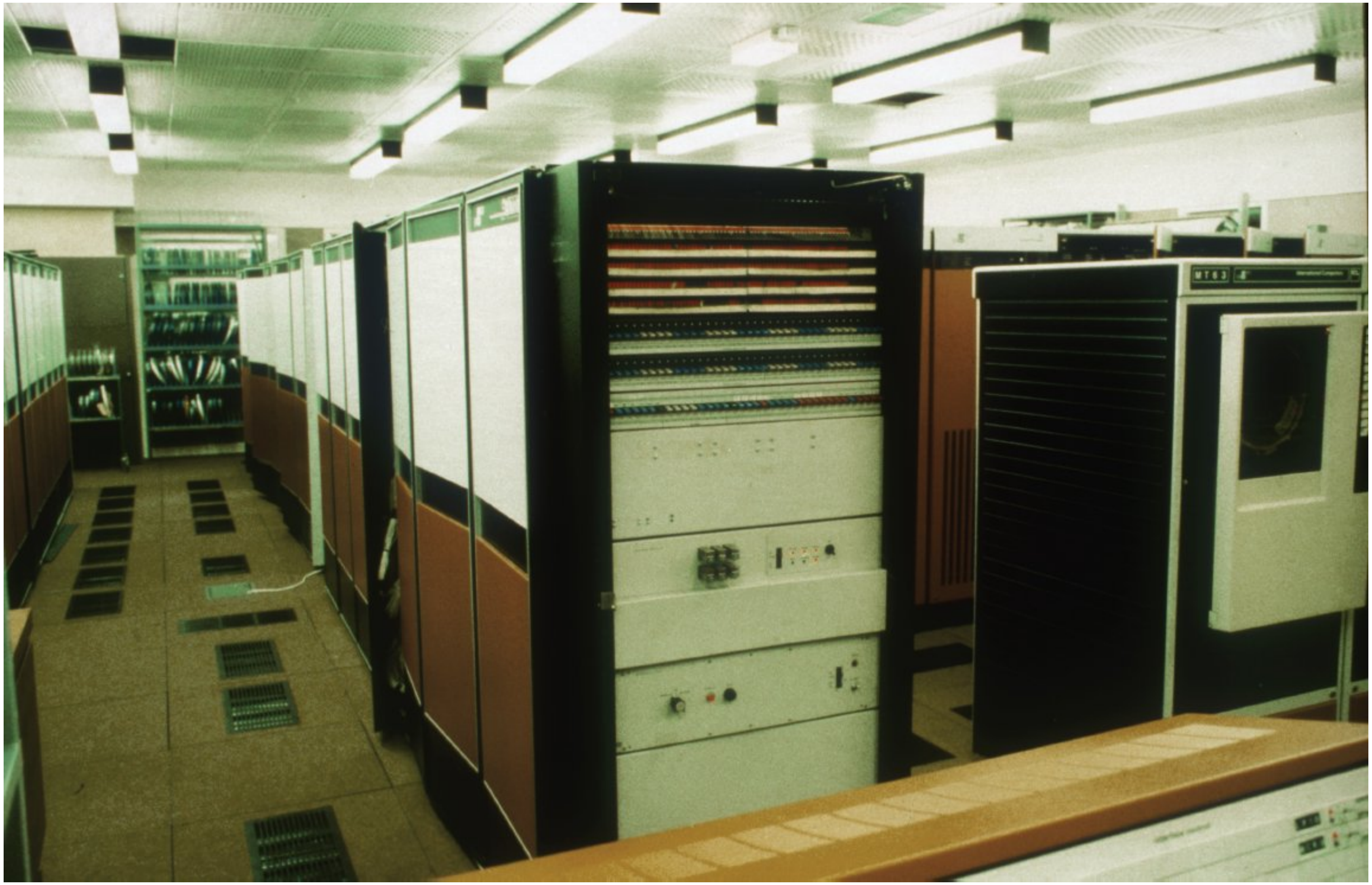
Performance Summary

- Useful definitions
  - Speed-up
  - Efficiency
- Amdahl's Law – *"the performance improvement to be gained by parallelisation is limited by the proportion of the code which is serial"*
- Gustafson's Law – to maintain constant efficiency we need to scale the problem size with the number of CPUs.

Parallel Computers at Edinburgh

- 1981 ICL DAP (SIMD; 4K processors)
- 1986 Meiko T800 CS (MIMD-DM; 400 processors)
- 1988 AMT DAP608 (SIMD; 1K processors)
- 1990 Meiko i860 CS (MIMD-DM; 64 processors)
- 1991 TMC CM-200 (SIMD: 16K processors)
- 1992 Meiko i860 CS (MIMD-DM; 16 processors)
- 1994 Cray T3D (MIMD-NUMA; 512 processors), Cray Y-MP (Vector)
- 1995 Meiko CS-2 (MIMD-DM)
- 1996 Cray J90 (Vector)
- 1997 Cray T3E (MIMD-NUMA; 344 processors)
- 1998 Hitachi SR2201 (MIMD-DM)
- 2000 Sun UltraSPARC III Cluster (SMP Cluster; 66 processors)
- 2001 Sun Fire 15K (MIMD-SMP; 52 processors)
- 2002 IBM p690 cluster (SMP cluster; 1280 processors)
- 2004 QCDOC (MIMD-DM; ~14,000 processors)
- 2005 IBM BlueGene/L (MIMD-DM; 2048 processors)
- 2006 IBM p575 cluster (SMP cluster; 2560 processors)
- 2007 Cray XE6 (MIMD-DM; 90,112 cores)
- 2013 Cray XC30 (MIMD-DM; 118,080 cores)

ICL DAPs



ICL DAPs

- Facts and Figures
  - *Lifetime:* 1981/1982--1988
  - *Processors:* 4096 single bit processors (in each of 2 systems)
  - *Peak Performance:* 0.03 GFlops
  - *Architecture:* SIMD
  - *Memory:* 2 MBytes
  - *Programming:* Data Parallel (DAP Fortran)
- Notes
  - one of the earliest production parallel computers
  - ICL made ~10 before AMT took the DAP technology forward
  - EPCC had a significant upgrade in 1988 to AMT DAP (1024 processors with 4 MBytes and a peak of ~60 Mflops)

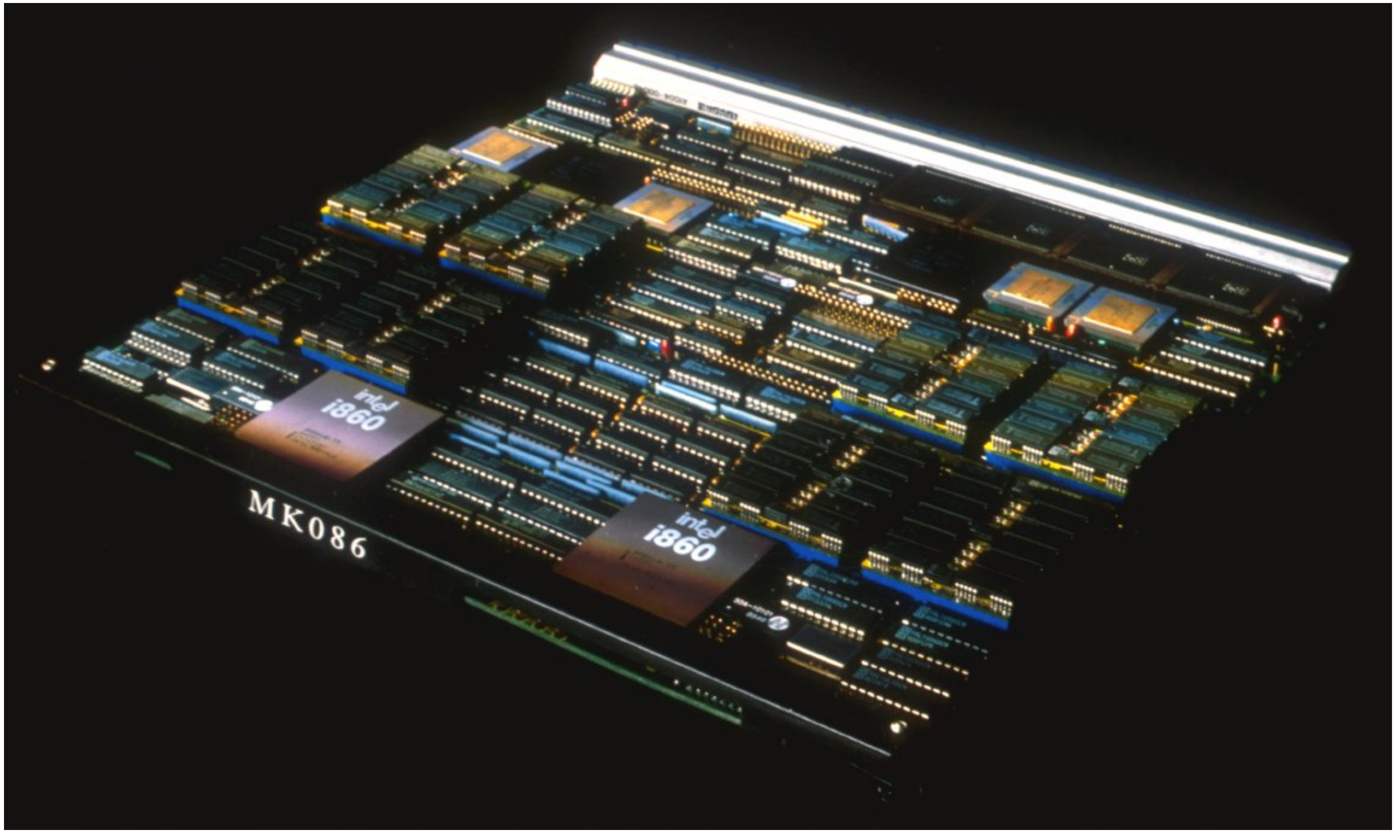
Meiko CS-1

Meiko CS-1

- Facts and Figures
  - *Lifetime:* 1986--1994
  - *Processors:* 400 x T800 Transputers
  - *Peak Performance:* 0.4 GFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 400 MBytes
  - *Programming:* OCCAM special purpose language/OS

- Notes
  - T800 was first processor to have a peak of 1 MFlops and also had built-in support for passing messages
  - focus for *Edinburgh Concurrent Supercomputer Project* which was pre-cursor to EPCC

Meiko i860

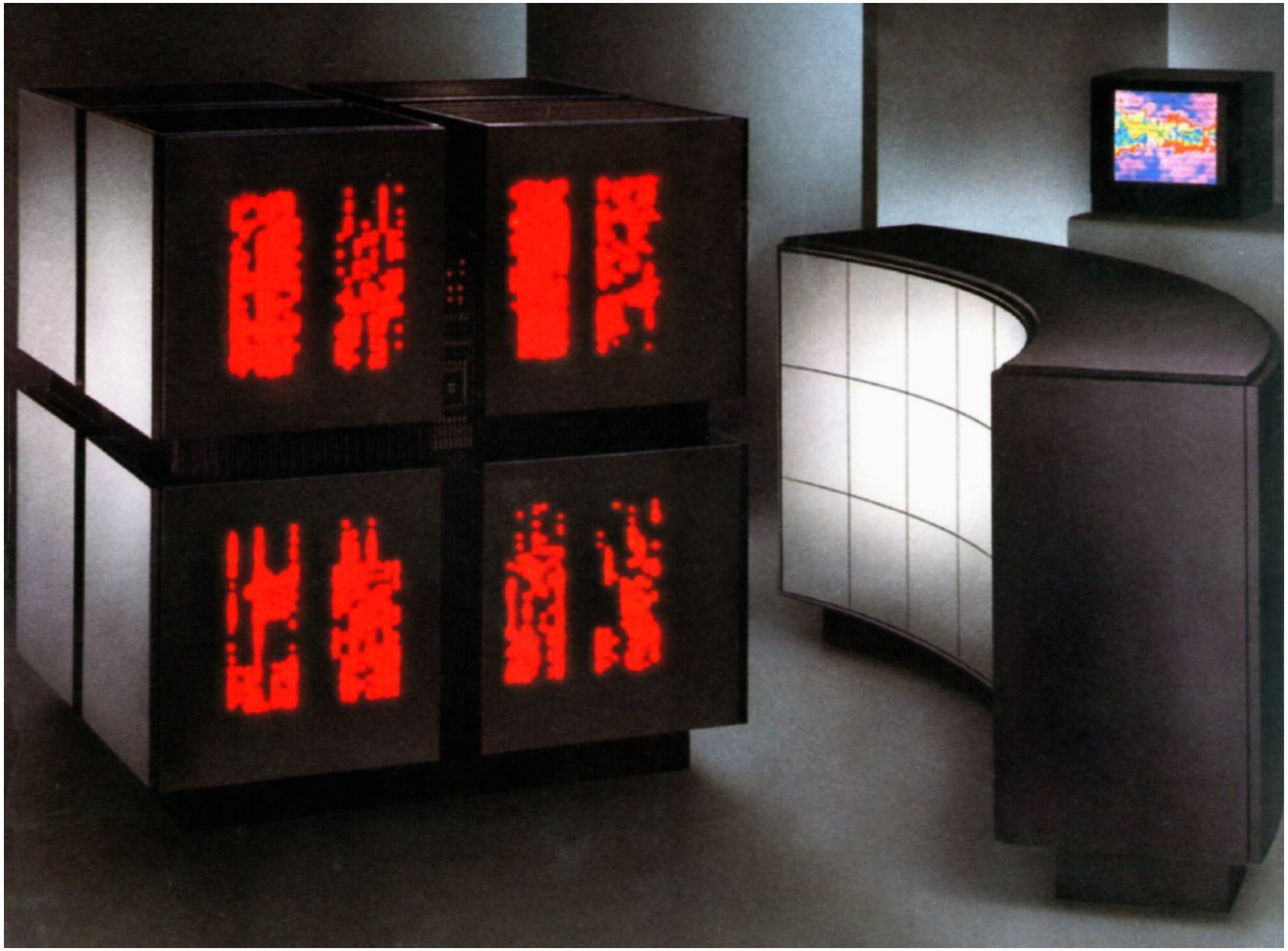


Meiko i860

- Facts and Figures
  - *Lifetime:* 1990--1995
  - *Processors:* 64 x 80 MHz i860 (+ T800s for communication)
  - *Peak Performance:* 5.1 GFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 1 Gbyte
  - *Programming:* Message Passing (CSTools)

- Notes
  - split between QCD and Materials Grand Challenges
  - QCD code sustained more than 1 GFlop making it one of the fastest applications codes in the world!

TMC CM-200

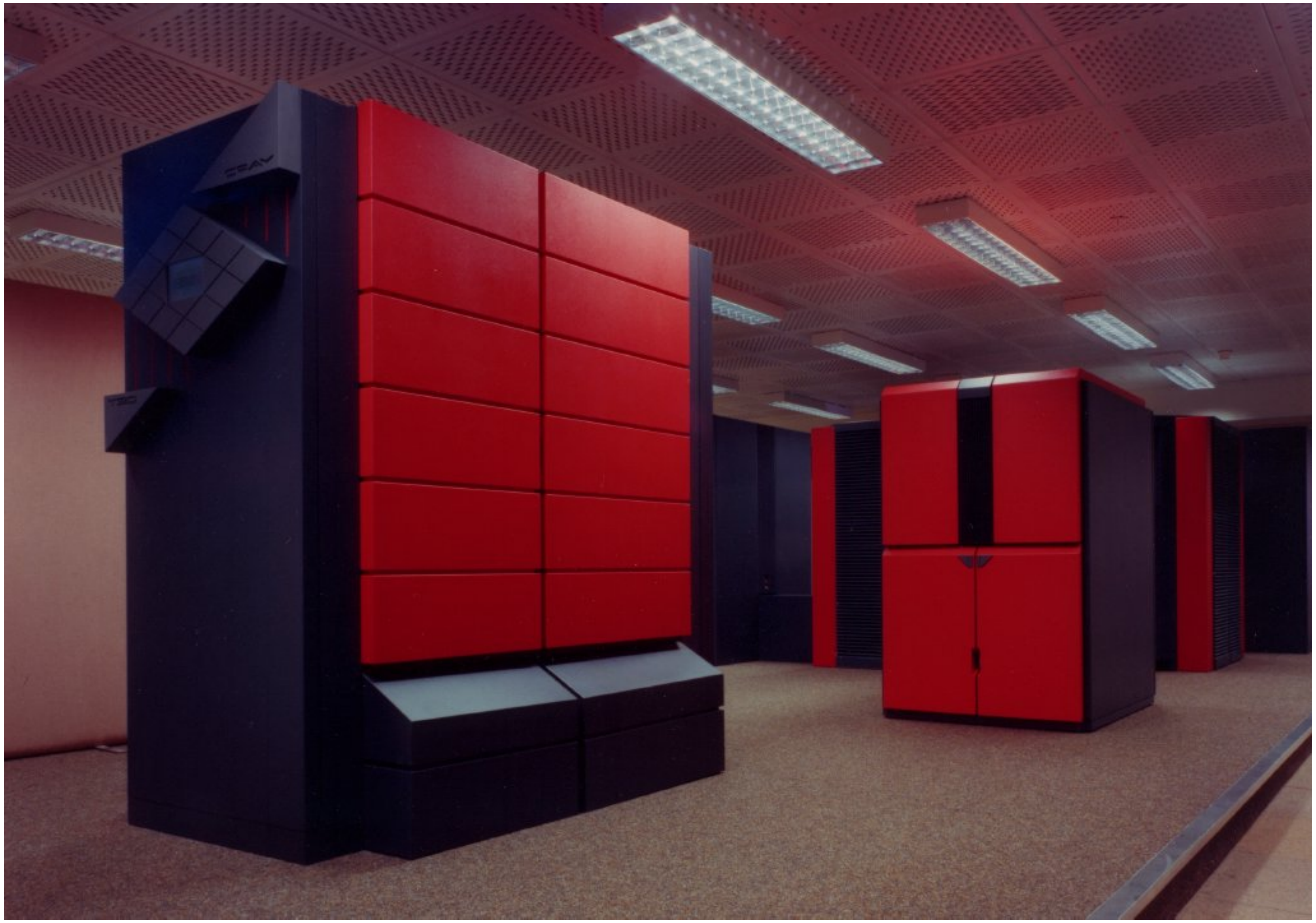


TMC CM-200

- Facts and Figures
  - *Lifetime:* 1991--1996
  - *Processors:* 16,584 single bit processors + 512 FPUs
  - *Peak Performance:* 5 GFlops
  - *Architecture:* SIMD
  - *Memory:* 512 MBytes
  - *Programming:* Data Parallel (CM Fortran, C*)

- Notes
  - largest SIMD machine in Europe
  - state-of-the-art Data Vault with 10 GBytes of storage

Cray T3D



Cray T3D

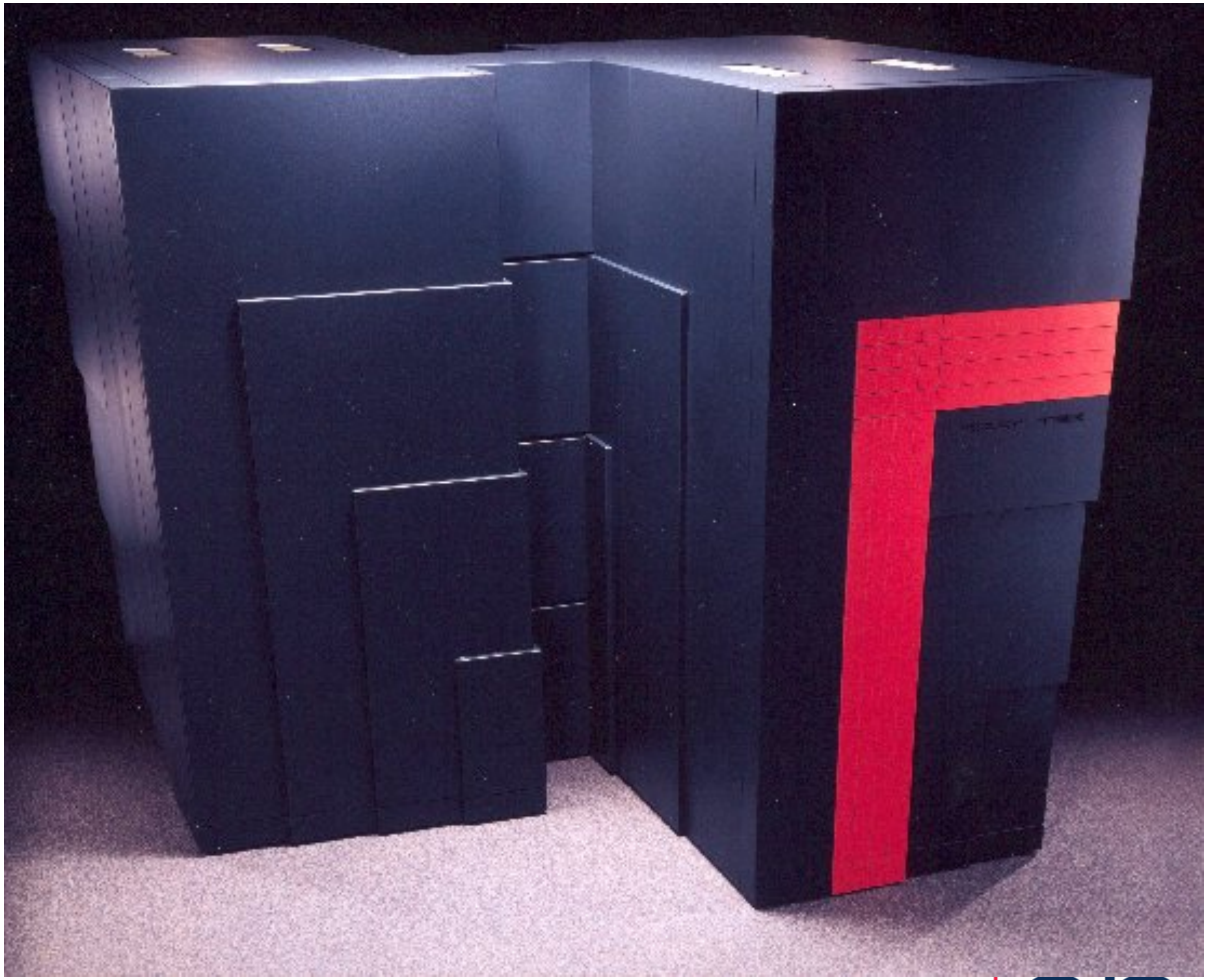
- Facts and Figures
  - *Lifetime:* 1994-1999
  - *Processors:* 512 x 150 MHz EV5 Alphas
  - *Peak Performance:* 76 GFlops
  - *Architecture:* MIMD-NUMA
  - *Memory:* 32 GBytes
  - *Programming:* Message Passing (PVM/MPI), Work Sharing, Data Parallel (CRAFT)

- Notes
  - first UK national parallel computing service
  - at various times, this was the largest T3D in Europe
  - choice of programming paradigms

Cray J90

- Facts and Figures
  - *Lifetime:* 1996--2002
  - *Processors:* 10 x 100 MHz Vector
  - *Peak Performance:* 2 GFlops
  - *Architecture:* MIMD-SMP/MISD?
  - *Memory:* 2 GBytes
  - *Programming:* serial/vector
- Notes
  - EPCC primarily had vector facilities in support of the Cray HPC systems

Cray T3E



Cray T3E

- Facts and Figures
  - *Lifetime:* 1997--2002
  - *Processors:* 344 x 450 MHz EV56 Alphas
  - *Peak Performance:* 310 GFlops
  - *Architecture:* MIMD-NUMA
  - *Memory:* ~40 GBytes
  - *Programming:* Message Passing (MPI), Data Parallel (CRAFT)

- Notes
  - supported multiple services for various communities
  - different processors had 64, 128 or 256 MBytes of memory

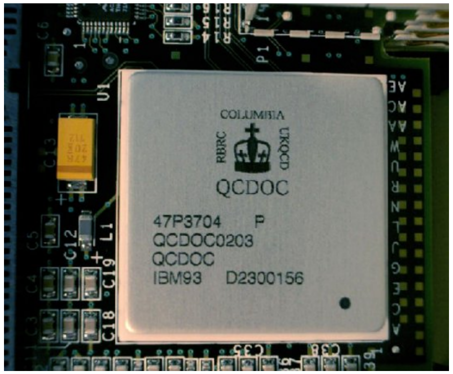
IBM p690 Cluster (HPCx Phase 1)

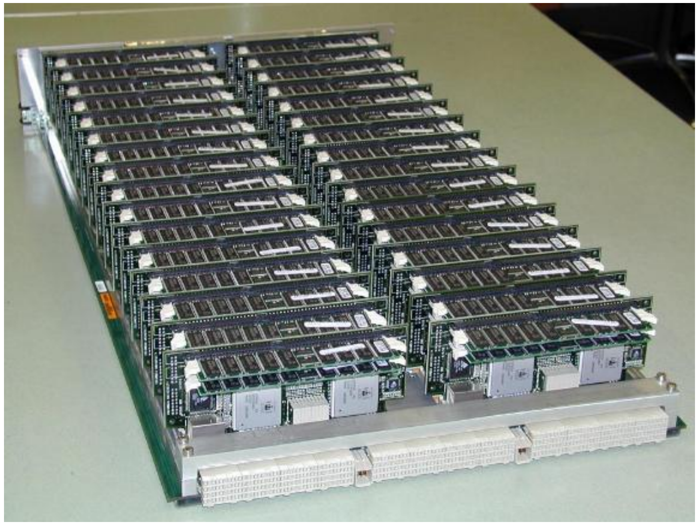
IBM p690 Cluster (HPCx Phase 1)

- Facts and Figures
  - *Lifetime:* 12/2002--7/2004
  - *Processors:* 1280 x 1.3 GHz Power 4s
  - *Peak Performance:* 6.7 TFlops
  - *Architecture:* SMP cluster
  - *Memory:* 1280 GBytes
  - *Programming:* Message Passing (MPI), Mixed Mode (OpenMP+MPI)

- Notes
  - UK national HPC service run by UoE/EPCC, DL and IBM
  - EPCC lead partner, although system is located at DL
  - focus on capability computing
  - upgrades to Power 4+ (2004) and then Power 5 (2005/06)

QCDoC





- **Q**uantum **C**hromo**D**ynamics **on** a **C**hip

QCDoC

- Facts and Figures
  - *Lifetime:* 10/2004 --
  - *Processors:* >14,000 x 400 MHz special-purpose chips
  - *Peak Performance:* ~11 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* ~1750 GBytes
  - *Programming:* Non-standard Message Passing (nearest-neighbour communications plus collectives)
- Notes
  - multiple systems (largest had 12K processors)
  - designed by IBM, University of Edinburgh and Columbia
  - QCD sustains up to 4 TFlops

IBM BlueGene (Blue Sky)

IBM BlueGene (Blue Sky)

- Facts and Figures
  - *Lifetime:* 1/2005-1/2018
  - *Processors:* 2048 x 700 MHz PowerPCs
  - *Peak Performance:* 5.6 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 512 GBytes
  - *Programming:* Message Passing (MPI)

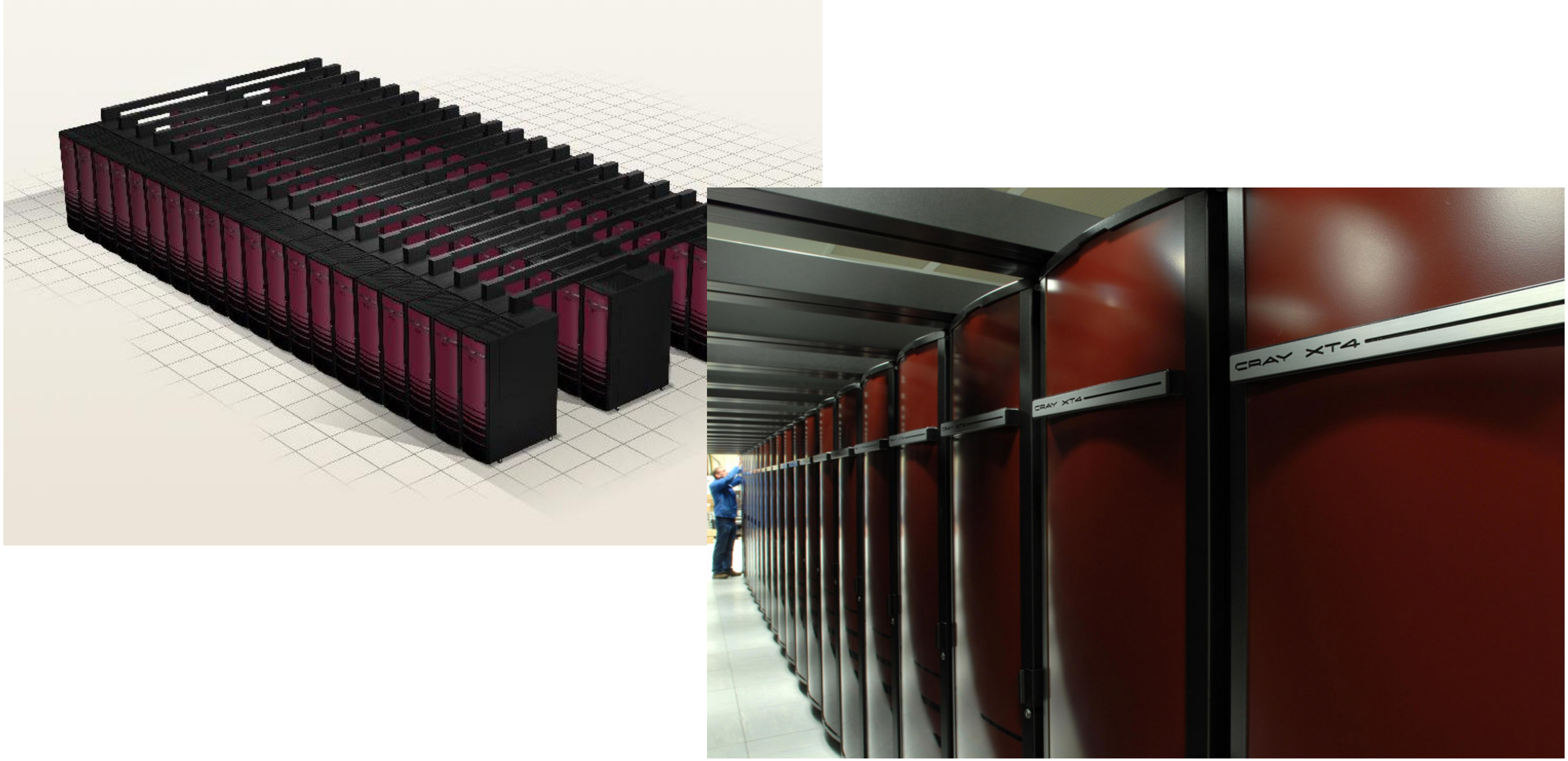
- Notes
  - first BlueGene system in Europe
  - low power requirements and high density of processors
  - capable of scaling to extremely large systems
    - many BlueGene systems of >100 TF

IBM p575 Cluster (HPCx Phase 3)

IBM p575 Cluster (HPCx Phase 3)

- Facts and Figures
  - *Lifetime:* 7/2006—2010
  - *Processors:* 2560 x 1.5 GHz Power 5s
  - *Peak Performance:* 15.3 TFlops
  - *Architecture:* SMP cluster
  - *Memory:* 5120 GBytes
  - *Programming:* Message Passing (MPI), Mixed Mode (OpenMP+MPI)
- Notes
  - double the memory per processor of Phase 1/2
  - significantly improved interconnect
  - focussed on Complementary Capability Computing
  - closes in January 2010

Cray XT4 (HECToR Phase 1)



Cray XT4 (HECToR Phase 1)

- Facts and Figures
  - *Lifetime:* 9/2007— 6/2009
  - *Processors:* 5664 x 2.8 GHz dual-core Opterons
    - i.e. 11,328 cores
  - *Peak Performance:* 63.4 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 34 TBytes
  - *Programming:* Message Passing (MPI)

- Notes
  - UK's previous national HPC facility
  - regular upgrades to 2013+
  - initially one of Top 20 systems worldwide
  - supplemented by Cray X2 (BlackWidow) vector system

Cray XT5 (HECToR Phase 2a)

- Facts and Figures
  - *Lifetime:* 6/2009 — 06/2010
  - *Processors:* 5664 x 2.3 GHz quad-core Opterons
    - i.e. 22,656 cores
  - *Peak Performance:* 208 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 45.3 TBytes
  - *Programming:* Message Passing (MPI)

- Notes
  - reduced in size from 60 to 33 cabinets in 2Q10
    - to allow for Phase 2b

Cray XE6 (HECToR Phase 2b)

- Facts and Figures
  - *Lifetime:* 6/2010 — 12/2011
  - *Processors:* 3712 x 2.1 GHz 12-core (Magny-Cours) Opterons
    - i.e. 44,544 cores
  - *Peak Performance:* 374 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 59.4 TBytes
  - *Programming:* MPI

- Notes
  - 20 XT6 cabinets
  - Gemini interconnect

Cray XE6 (HECToR Phase 3)

- Facts and Figures
  - *Lifetime:* 12/2011 — 2014
  - *Processors:* 5632 x 2.3GHz 16-core (Interlagos) Opterons
    - i.e. 90,112 cores
  - *Peak Performance:* >800 TFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 90 TBytes
  - *Programming:* MPI

- Notes
  - 30 XE6 cabinets
  - Gemini interconnect

Cray XC30 (ARCHER)

- Facts and Figures
  - *Lifetime:* 11/2013 — 2020
  - *Processors:* 9840 x 2.7 GHz 12-core Intel Xeon
    - i.e. 118,080 cores
  - *Peak Performance:* >2.5 PFlops
  - *Architecture:* MIMD-DM
  - *Memory:* 400 TBytes
  - *Programming:* MPI
- Notes
  - 26 XC30 cabinets
  - Aries interconnect

Experiences of EPCC

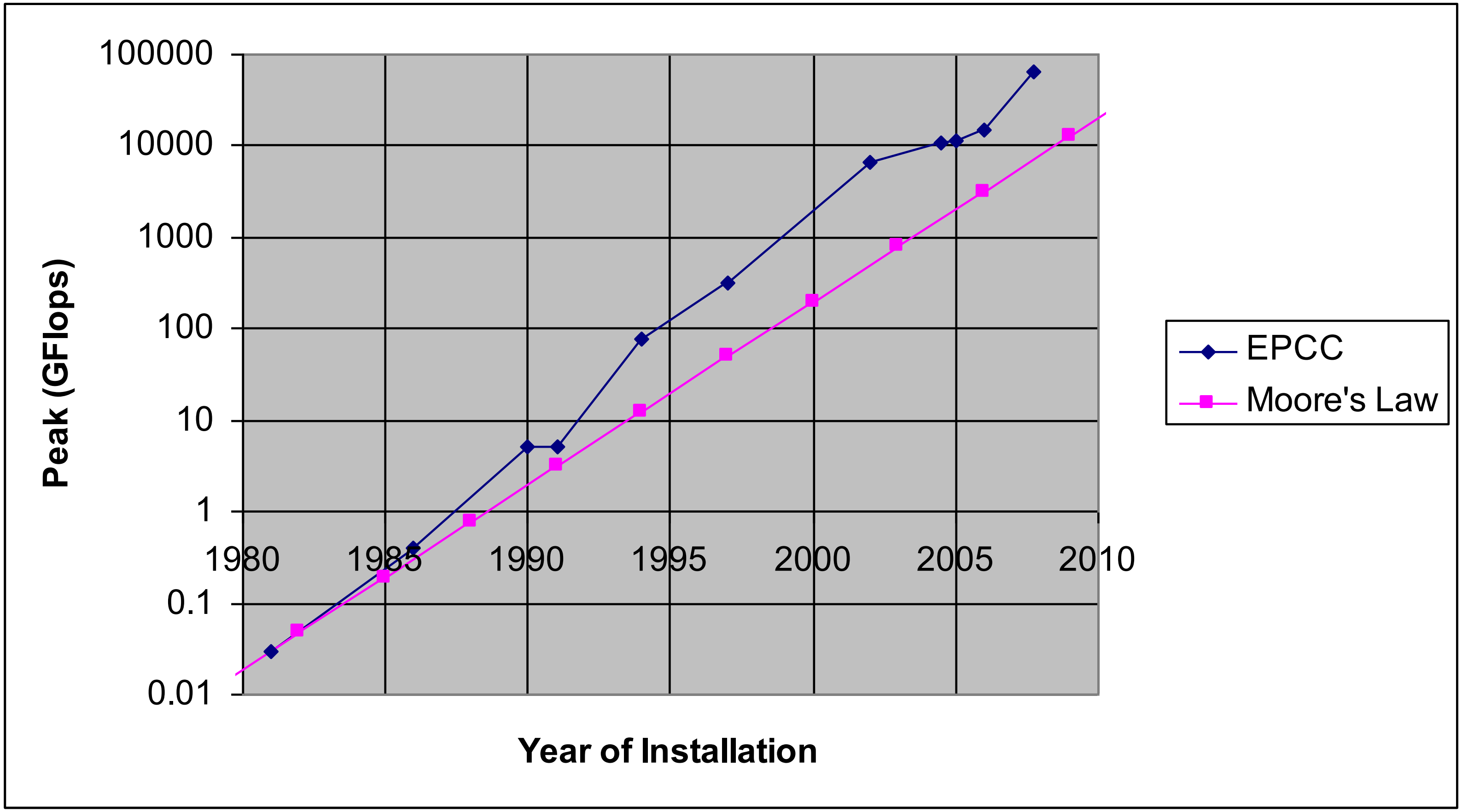
- New generation of machines every 3-4 years

- Each generation providing
  - significantly greater compute power and larger memories
  - better tools and more stable programming environment
  - … but number of processors has increased only modestly

- Useful lifetime of machines around 5 years

- (Nearly) missed out on vector architectures

Peak Performance



Peak Performance

- EPCC has managed to stay ahead of Moore's Law for the last 25 years!

- Most of the major systems were initially in the top 20 or so worldwide
  - see **www.top500.org**

- GFlops in 1990
- TFlops in 2002
- PFlops in 2012 - (DiRAC Bluegene/Q)
- Eflops by 2022-2024

The Future

- The end of Moore’s Law?
  - rise of multi-core processors

- HPC increasingly mainstream
  - animation for Shrek, Lord of the Rings, etc.
  - PlayStation 3 based on multi-core Cell architecture
  - Multicore CPUs and GPGPUs
  - Big data, big cloud, etc…

- Exciting range of HPC-oriented architectures
  - Massively parallel (BlueGene, XC)
  - Accelerators (Xeon Phi, GPGPU)
  - Shared memory clusters (SGI UV)
  - Low power (ARM)
  - Big data (Hadoop, Spark)

Summary

- Parallel computing started as a range of weird and wacky architectures
  - with their own unique languages, OS, tools, etc.
  - each required substantial investment of time to port to
- Standardisation of programming paradigms was vital for parallel computing to mature
  - MPI, HPF, OpenMP,…
- Edinburgh/EPCC has been at the forefront of parallel computing and HPC for 30 years
- The future appears extremely exciting for HPC, computational science and EPCC