Past, present, and future of HPC at ECMWF

21st ECMWF workshop on high performance computing in meteorology

Mike Hawkins

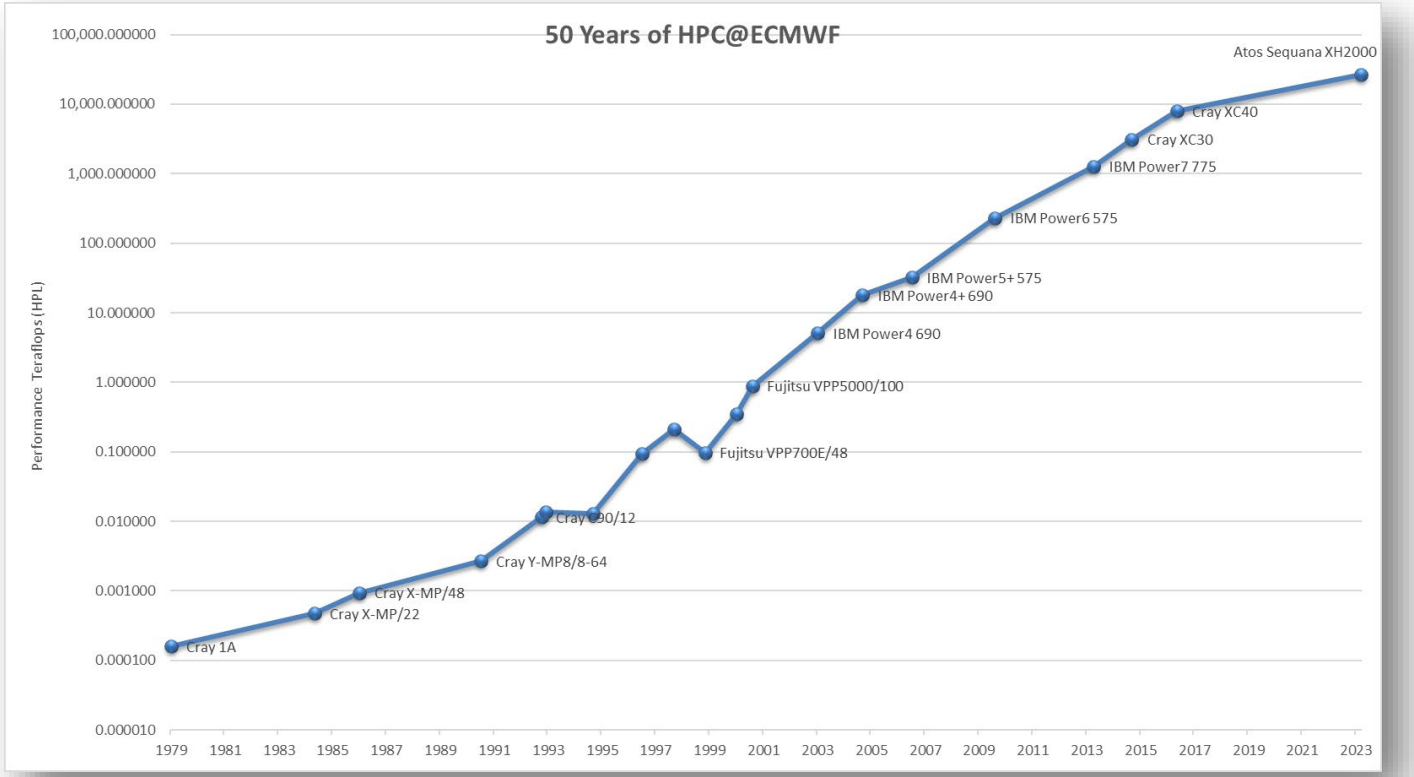
ECMWF

Michael.Hawkins@ecmwf.int

ECMWF

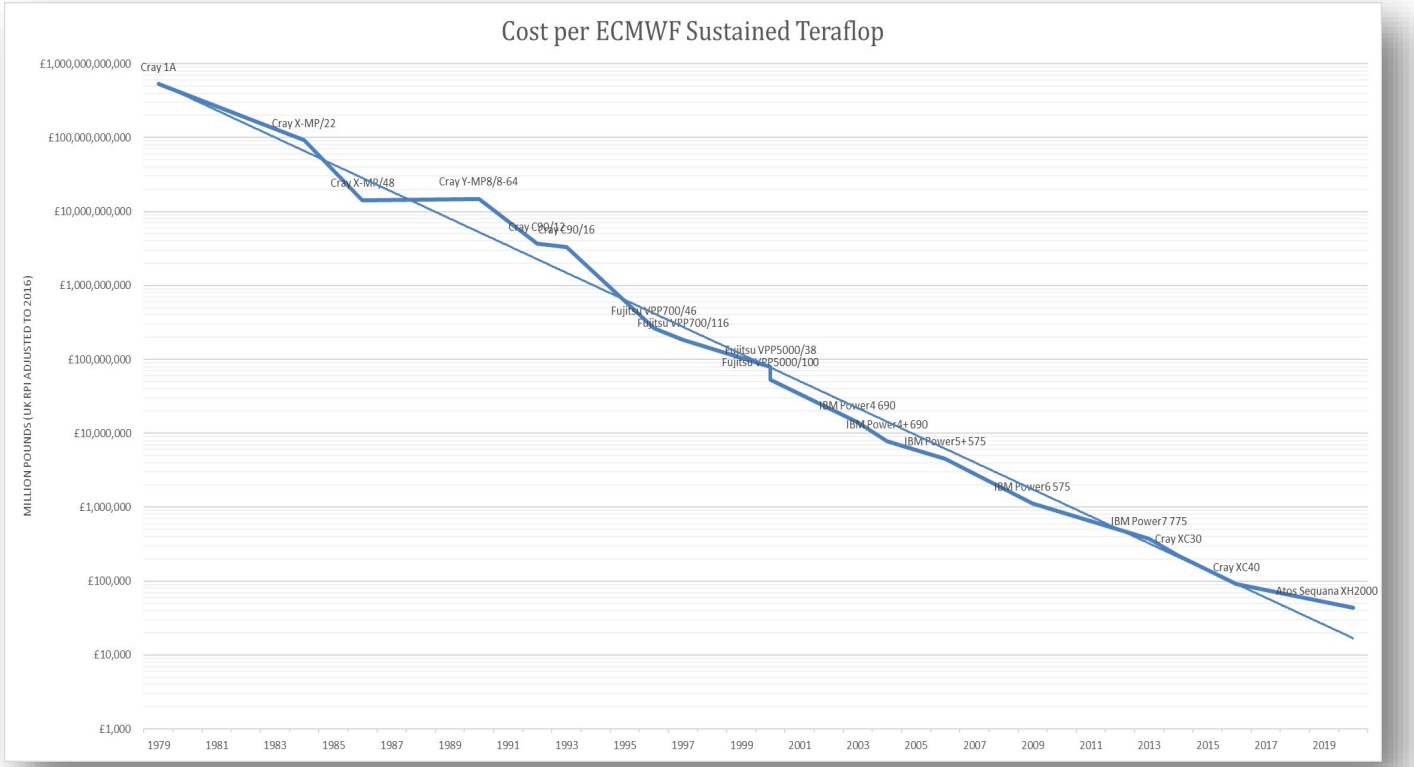
ECMWF: 2025

- 1978 to 2025
  - Performance increased from 160 million floating point operations per second to 26 quadrillion floating-point operations per second **(HPL x166M, ECMWF x30M)**
  - Data increased from 500 gigabytes to 1.5 exabytes **(x30M)**
  - Power consumption rose from 140 kW to 4.5 MW **(x32)**



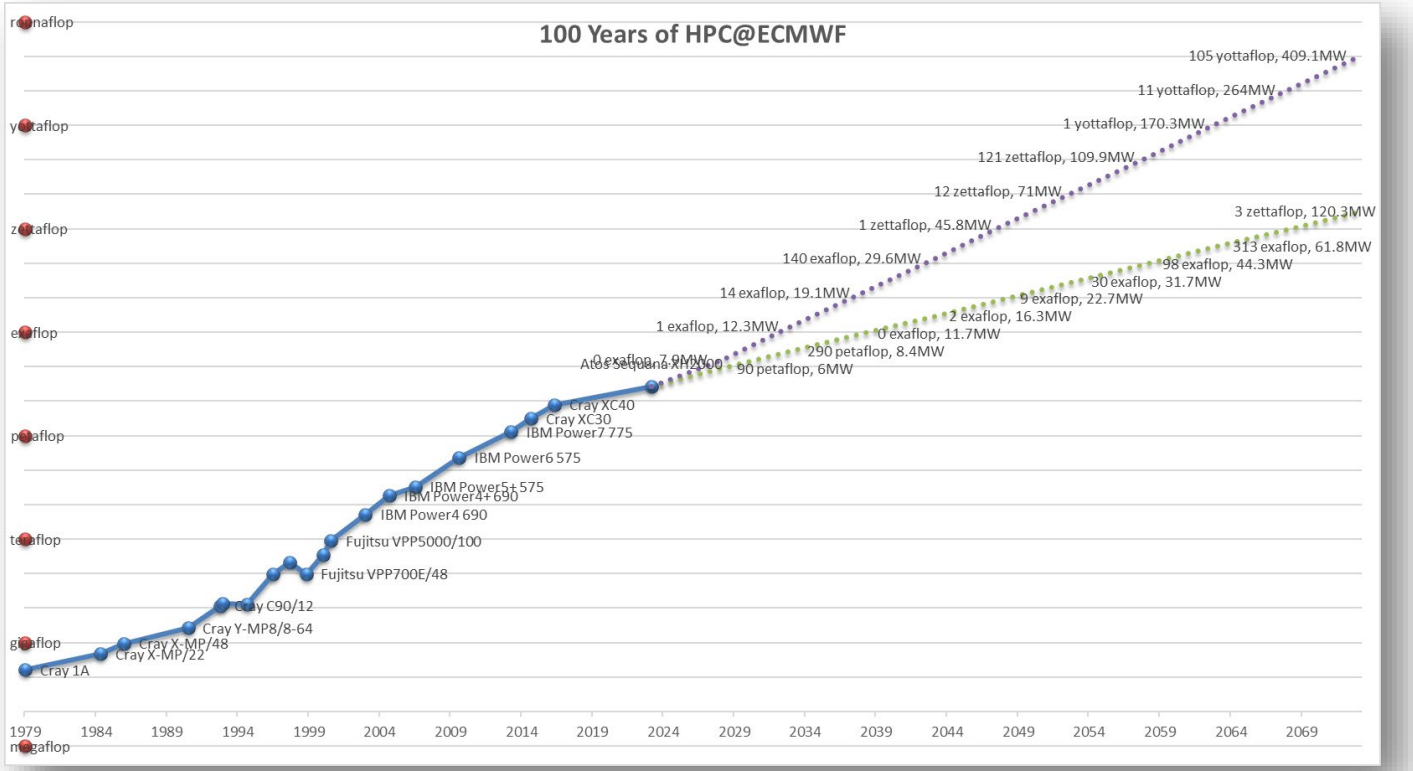
ECMWF: 2025

- 1978 to 2025
  - Performance increased from 160 million floating point operations per second to 26 quadrillion floating-point operations per second **(HPL x166M, ECMWF x30M)**
  - Data increased from 500 gigabytes to 1.5 exabytes **(x30M)**
  - Power consumption rose from 140 kW to 4.5 MW **(x32)**

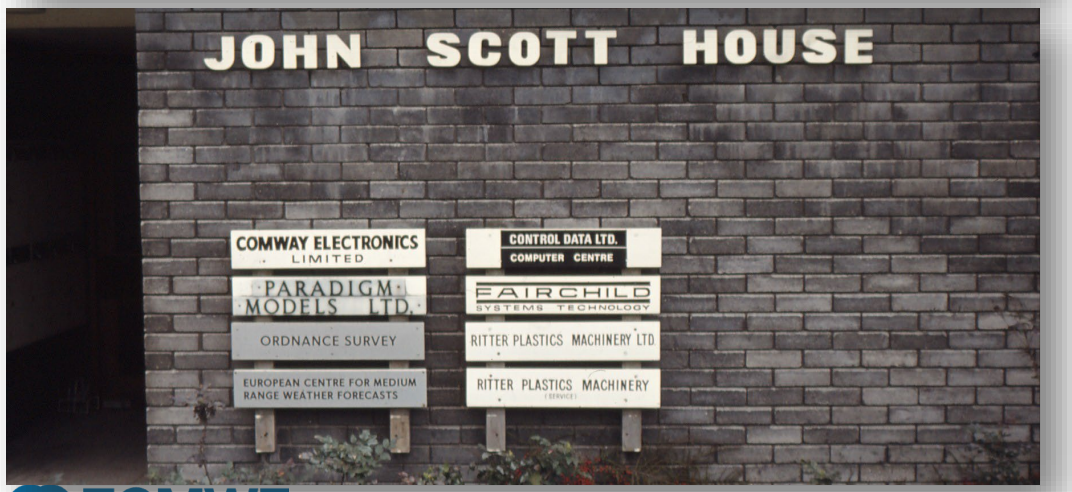


ECMWF: 2075

- 1978 to 2025
  - Performance increased from 160 million floating point operations per second to 26 quadrillion floating-point operations per second **(HPL x166M, ECMWF x30M)**
  - Data increased from 500 gigabytes to 1.5 exabytes **(x30M)**
  - Power consumption rose from 140 kW to 4.5 MW **(x32)**
- 2025 to 2075
  - Expected growth to between 3 zettaflops **(x121k)** to 105 yottaflops **(175M)**
  - Data increase projected to be between 1.6 yottabytes and 175 yottabytes (million exabytes)
  - Power usage projected to increase to between 120 to 400 MW
  - But iphone66 would have 0.5 exaflop of performance



The beginning - 1975



- No Shinfield Park data center, so time was rented on a CDC 6600 computer to run 200 jobs weekly.
- Only 40 hours of CPU time were available each week.
- CDC6600:
  - 60-bit processor @ 10 MHz
  - Up to 982 kilobytes
  - 2 MIPS
- Lacked the power for operational forecasting; a 10-day forecast would take 12 days to run.

Cray and the vector years

- The Cray-1, launched in 1976
- ECMWF used No.1 at Rutherford Laboratory
  - Courier service for magnetic tape
- System No.9 in 1978 in Shinfield Park
- Cray 1A:
  - 160 MFLOPS
  - 64-bit processor @ 80 MHz
  - 1,048,576 words of central memory (8MB)
  - Dedicated hardware for addition/subtraction and multiplication
  - separate pipelines for various instructions
  - pipeline parallelism for vector instructions
  - 8 vector registers, which held sixty-four 64-bit words
- Forecast time reduced from 12 days to 5 hours.

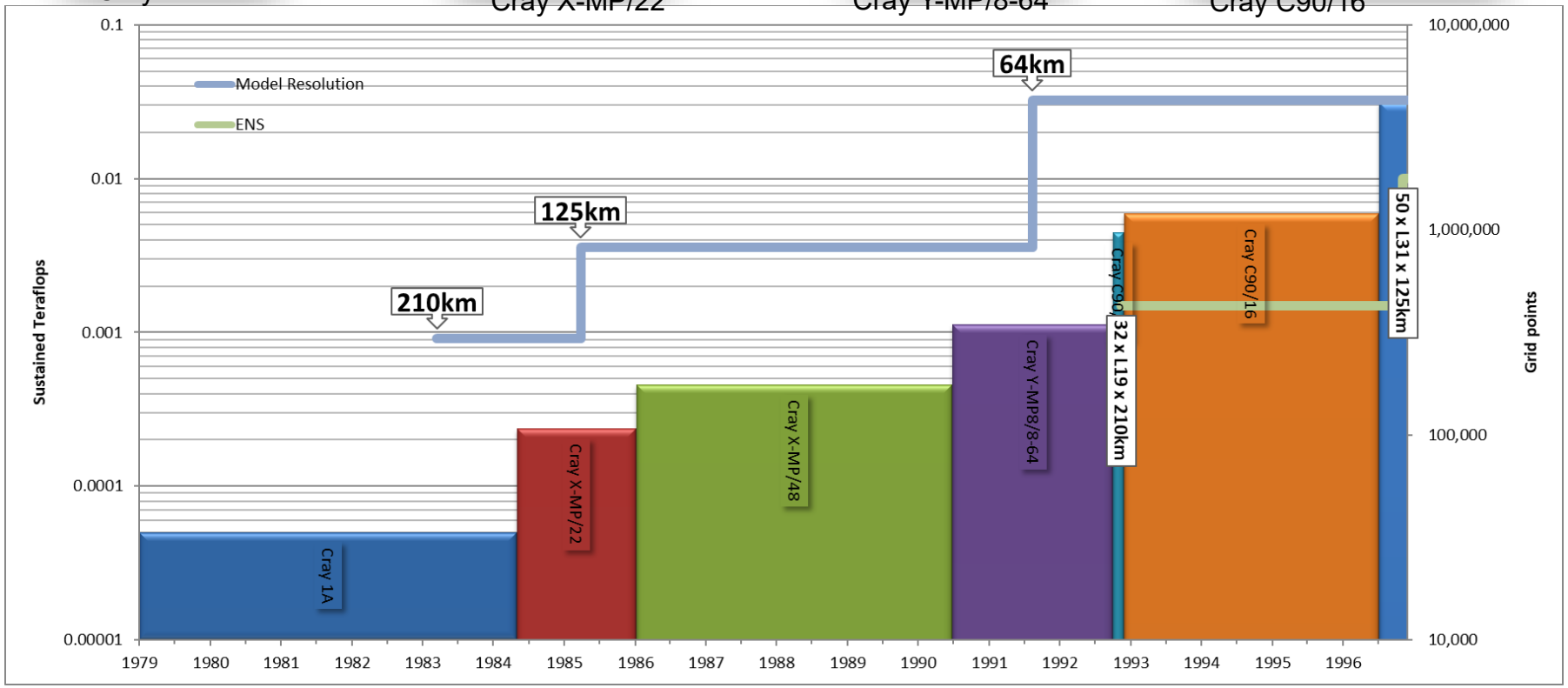
Megaflops to Gigaflops

Cray 1A

Cray X-MP/22

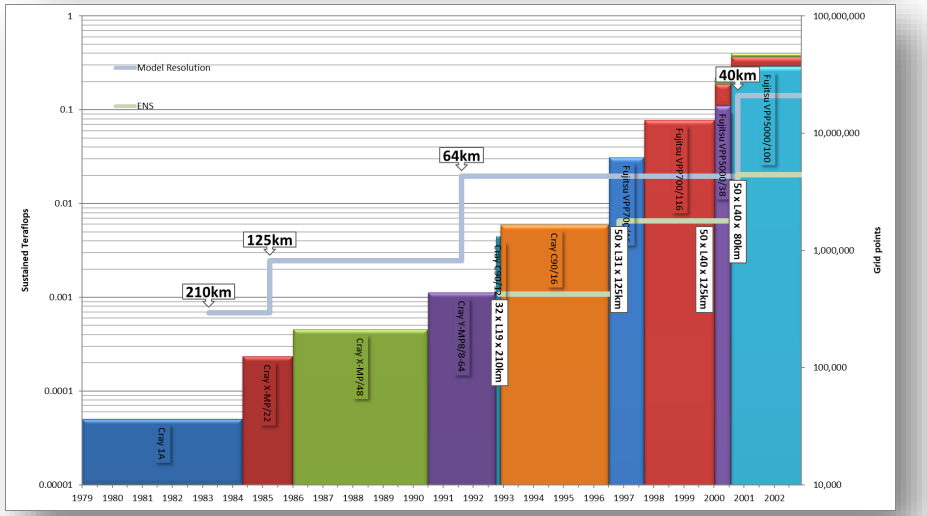
Cray Y-MP/8-64

Cray C90/16



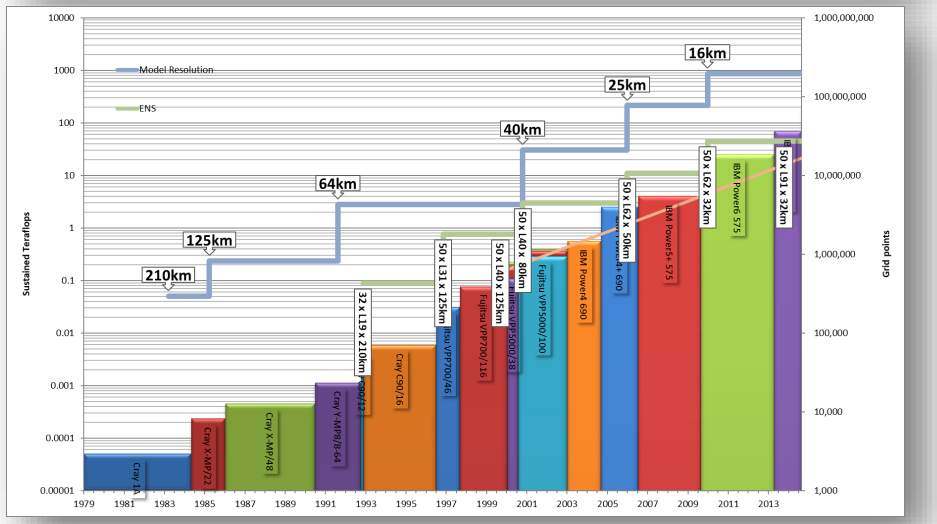
1994-2003: Fujitsu and the start of distributed memory parallel processing

- 1994 – Cray T3D
  - 128 Alpha microprocessors
    - 128 Mbytes of memory.
  - 3D torus.
- 1996 - Fujitsu VPP700
  - first operational distributed memory vector-parallel machine
  - 39 processors, 6 for I/O, and 1 for system tasks,
  - each with 2 GB local memory
  - Over 600 times the performance of the Cray-1A.
  - 1088GB Disk
- 1999 – Fujitsu VPP5000
  - 38 processors. (100 PEs in 2000)
  - Each processor in the VPP5000 was capable of a peak performance of 9.6 Gigaflops, more than 4 times faster than those in the VPP700.
  - VPP5000 ~ 300 Gigaflops.



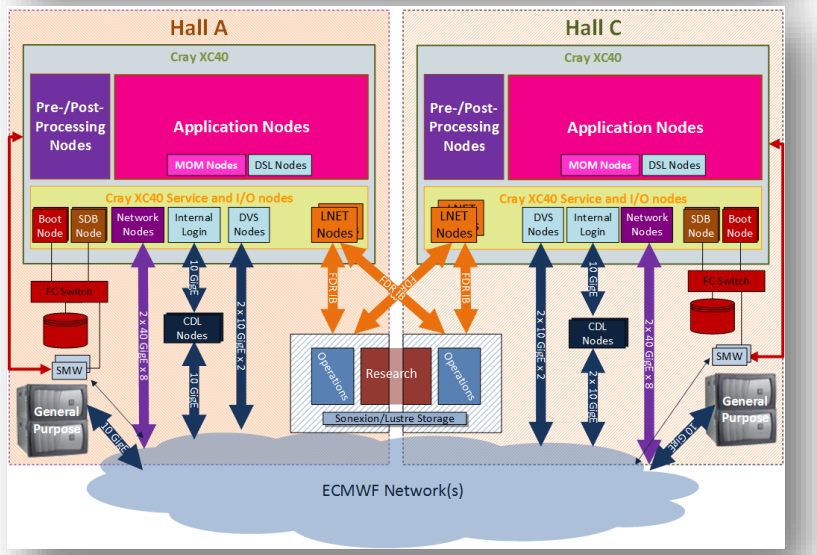
2003-2014: The IBM years – massively parallel

- Developing to cope with:
  - No Vector
  - Shared memory nodes with high performance interconnect
- 2003 - ECMWF's first massively parallel system
  - **2 Clusters**
  - IBM Power 4
    - 20 nodes, 32 processors each
    - 32/128 GiB memory
  - 10 TB of disk
- 2012: IBM Power7
  - IBM POWER7 8 cores, 3.836GHz
  - 4 Processors per node
  - 752 nodes
  - 64/256 GiB memory



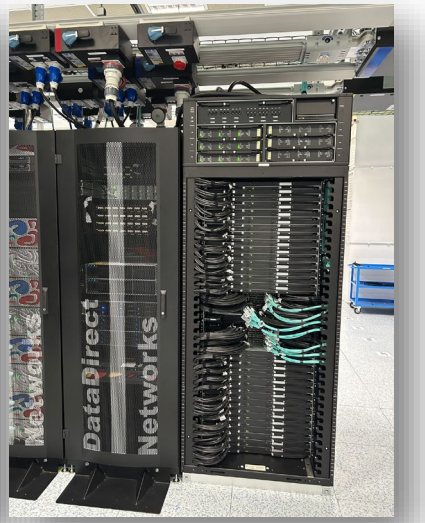
2014-2023: Linux, Lustre and lots and lots of commodity chips

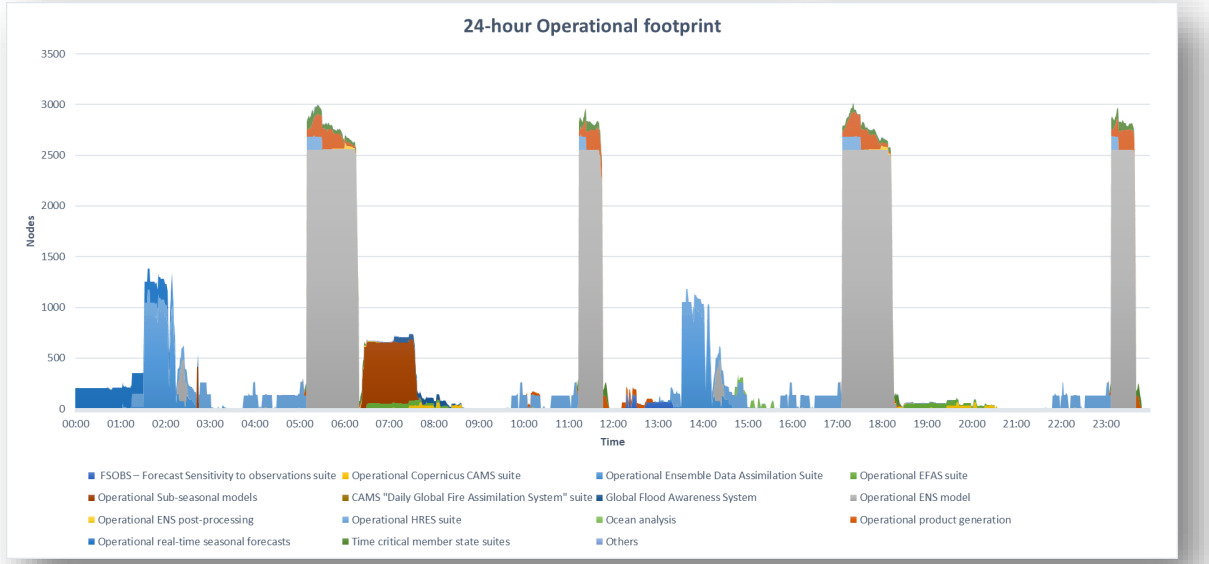
- Intel X86 architectures dominated the world
- Linux takes over from UNIX flavors
- Cray XC30
  - 168,000 processor cores
  - 2 clusters
  - Dragonfly topology for high performance network
  - 3505 nodes
    - 2 x 12-core Intel Ivy bridge Processors
    - 64GiB memory
- 2016 – In place upgrade to Intel Broadwell
  - 260,000 processors cores
  - Intel Xeon EP E5-2695 V4 “Broadwell” 120W 18-core
- Lustre storage:
  - Time Critical : 2 x 4.3PB
  - Research: 2 x 12PB



The present

- Atos BullSequana XH2000
  - 1 Million+ processor cores
- 4 clusters
  - 1920 nodes
  - 2 x AMD ‘Rome’ 7742 processors, 64-cores, 225W
  - 256 GiB
- GPU
  - 32 nodes NVIDIA A100
    - 4 x A100 per node
  - 30 Nodes Grace Hopper
    - 4 Superchips per node
- Lustre
  - 2 x Time Critical – short term SSD 1.9 PB
  - 2 x Time critical – medium term, 6.1PB
  - 6 x Research – 15 PB





Access: Card decks to terminals and beyond

- Users created extensive job decks using punch cards and submitted them to the computing center.
- By 1979, ECMWF had 20 alphanumeric visual display units and 4 graphical units in operation.
  - Several terminals even in staff offices!
- 1982, most staff members had a terminal in their office, resulting in a notable increase in productivity.
- 1990s: Desktops and workstations
- 2020s: Laptops and remote access
- Future: Extended reality?

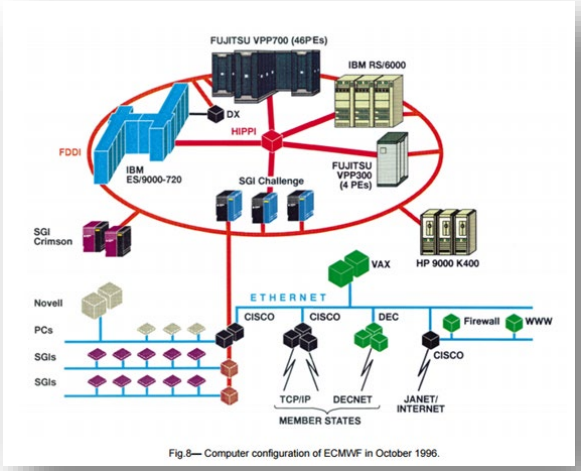
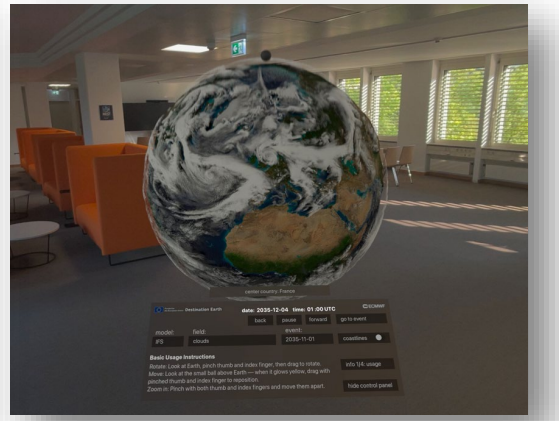
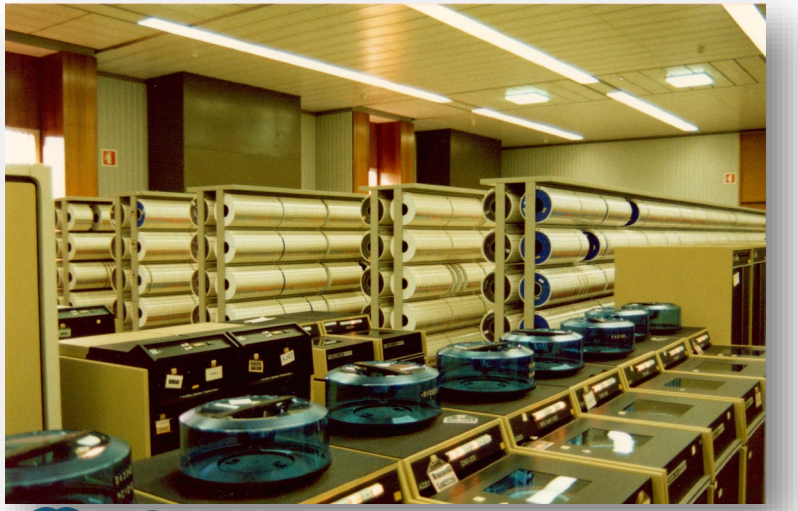


Fig.8— Computer configuration of ECMWF in October 1996.



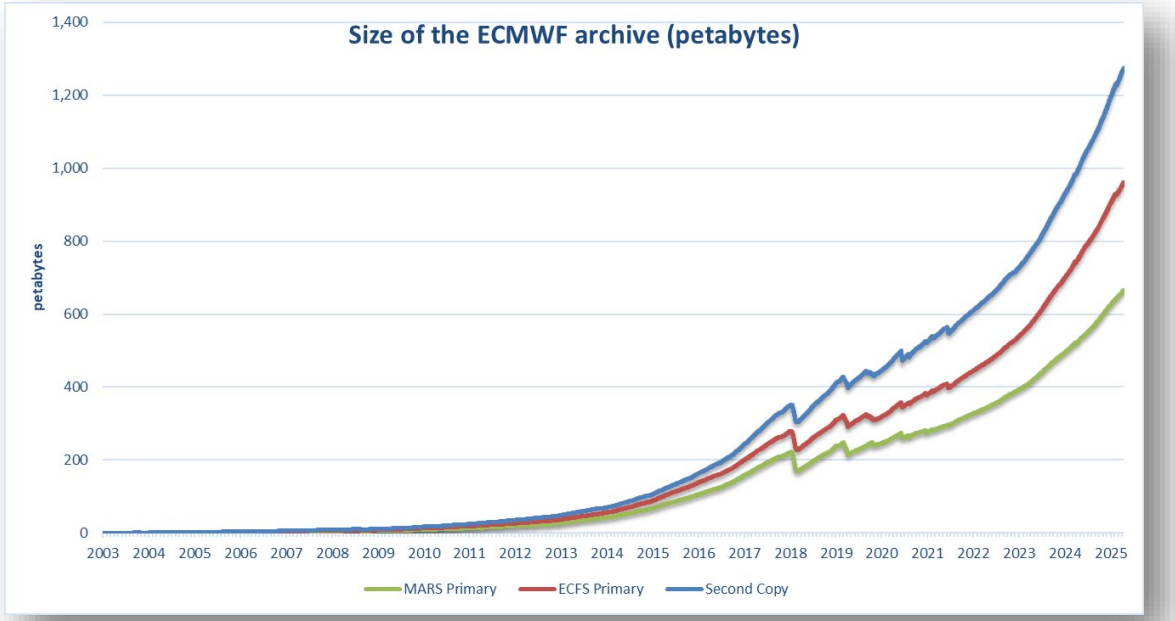
Data – Gigabytes to Exabytes

- Control Data 669-2 open-reel tape drives, each reel held around 40 MB
  - 1982 - 1,000 gigabytes of data, 15,000 tapes with 10 magnetic tape drives.
- 1984 – IBM single-reel cartridge tapes - by 1990 - 10 drives, 21,000 cartridge tapes
- 1991 - automated StorageTek 4400 robotic libraries
  - Up to 6,000 tape cartridges per silo - 200 MB to 1 TB – providing a 5,000-fold increase in capacity in the same physical space.
- 2009, SL8500 libraries from Sun Microsystems (later Oracle),
  - 10,000 tapes.64 tape drives in each library, 4 libraries, capacity grew from 1 TB to 5 TB to 8.5 TB.



Move to Bologna

- Moved to Italy:
  - 3 tape libraries,
  - 17 disk systems,
  - 150 servers,
  - 315 tape drives, 47,000 tapes,
  - Nine truckloads, 40,000 kilos
- Today:
  - ~ 1.3 exabytes of data
  - 31,000 primary, 27,000 secondary cartridges,
  - 14 tape libraries, 290 Linux servers, 34PiB disk,
  - 396 IBM 3592 tape drives, 60 LTO tape drive
  - In a typical day:
    - 18,500 tape mounts
    - 800 TB added , and 350 TB is retrieved.



Next Steps

- Next HPC procurement to be launched soon (https://www.ecmwf.int/en/about/suppliers)
- For:
  - A highly-available computational backbone for ECMWF's time-sensitive numerical weather prediction using **both physics and data-driven models.**
  - resources specifically optimized for the training of machine learning models.
  - Installation during 2027