ICES Foundation Biennial Workshop VII Panel: Challenges for HPC in AI dominated world

May 2024

2	128	**Isambard-AI phase 1** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom	34,272	7.42	117	68.835

Sadaf R Alam
University of Bristol
ICES Foundation Biennial Workshop VII
October 4, 2024

University of BRISTOL

A bit about myself

- 2022-today GB
  - University of Bristol-–Chief Technology Officer (CTO) at Bristol Centre for Supercomputing
- 2009-2022 CH
  - Swiss National Supercomputing Centre—various roles including CTO, head of operations & computer scientist
- 2004-2009 US
  - Oak Ridge National Laboratory (ORNL) Leadership Computing Facility liaison, computer scientist and postdoc
- 2000-2004 GB
  - University of Edinburgh PhD Researcher
- 1999-2000 GB
  - Analog Devices Software Engineer

International collaborations

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

EuroHPC
LUMI

FENIX
RESEARCH INFRASTRUCTURE

https://insight.ieeeusa.org/articles/the-art-of-mentoring/

University of BRISTOL

University of Bristol & Isambard Projects

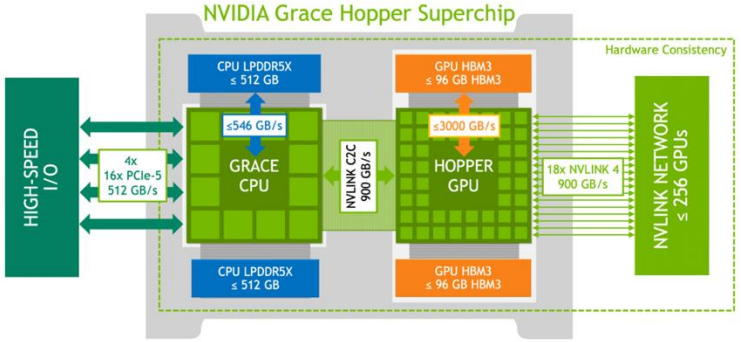
- The University of Bristol is a red brick Russell Group research university in Bristol, England. It received its royal charter in 1909, although it can trace its roots to a Merchant Venturers' school founded in 1595 and University College, Bristol, which had been in existence since 1876 [Wikipedia]
- The GW4 (Bath, Bristol, Cardiff and Exeter) Isambard project initially set out to prove that a new **ARM-based** technology was relevant to supercomputing since 2016
- **Bristol Centre for Supercomputing (BriCS)** has been recently formed for managing Isambard Digital Research Infrastructure (DRI) projects within the Faculty of Engineering

University of BRISTOL

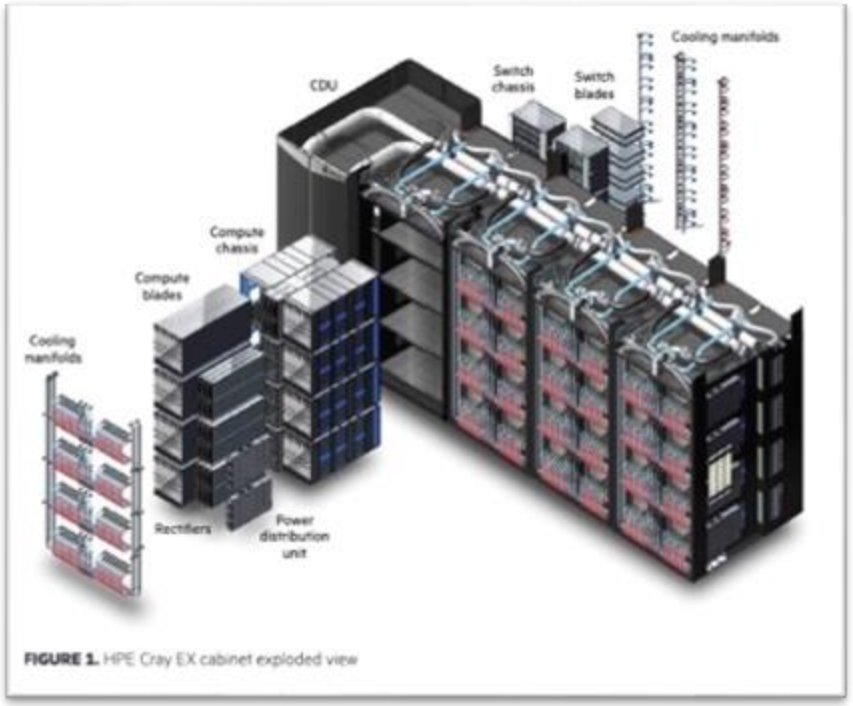
Panel Q: Addressing energy and power costs and requirements

Isambard-AI Sustainability Recipes using Modular Data Centre, DLC, Energy Efficient Compute, and GHG Costings

Modular Data Centre



ARM and GPU Computing



Direct Liquid Cooling

What are Scope 1, 2 and 3 carbon emissions?

The three scopes are a way of categorising the different types of greenhouse gas emissions created by a company, its suppliers and its customers.

nationalgrid

Green House Gas (GHG) Scope 1, 2 & 3 Emissions

University of BRISTOL

Isambard-AI vs Nvidia DGX SuperPOD

Isambard AI

- Space
  - 5280 GH200 in 12 HPE Cray EX compute cabinets
- Power
  - ~5MW inclusive of ecosystem
- TCO
  - CapEx ~£200M, TCO ~£300M over 5 years
- Software stack
  - Work in progress (see next slide)

DGX SuperPOD

- Space
  - 5,280 GH200 in ~330 compute cabinets
- Power
  - ~13 MW (40kW per H100 rack)
- TCO
  - CapEx >£400M, TCO >£600M over 5 years
- Software stack
  - Nvidia AI and ML optimised stack

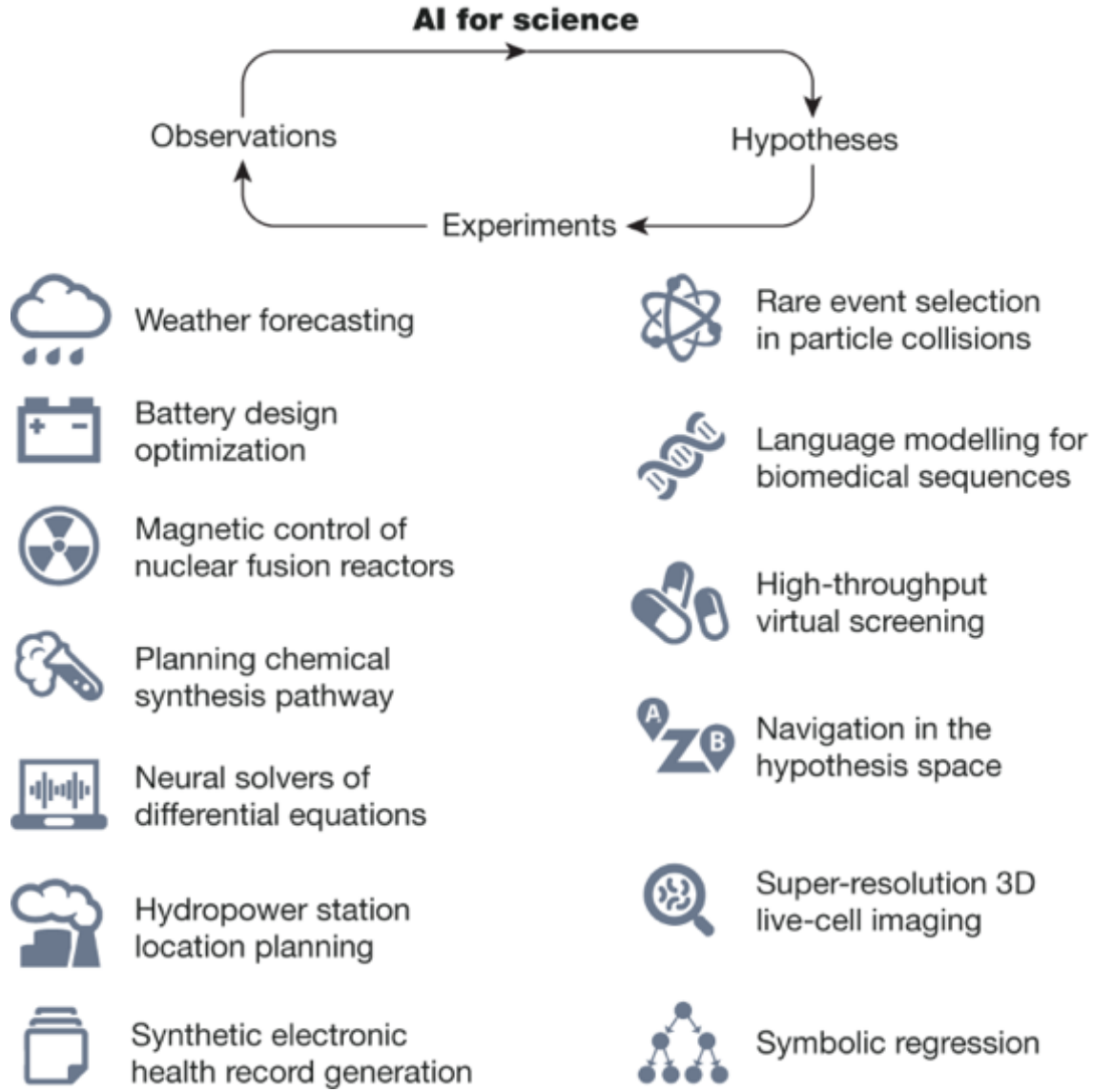
Disclaimer: information is based on publicly available data and ChatGPT responses

University of BRISTOL

Panel Q: Availability of adequate computing technology and of competent brain ware

AI for Science

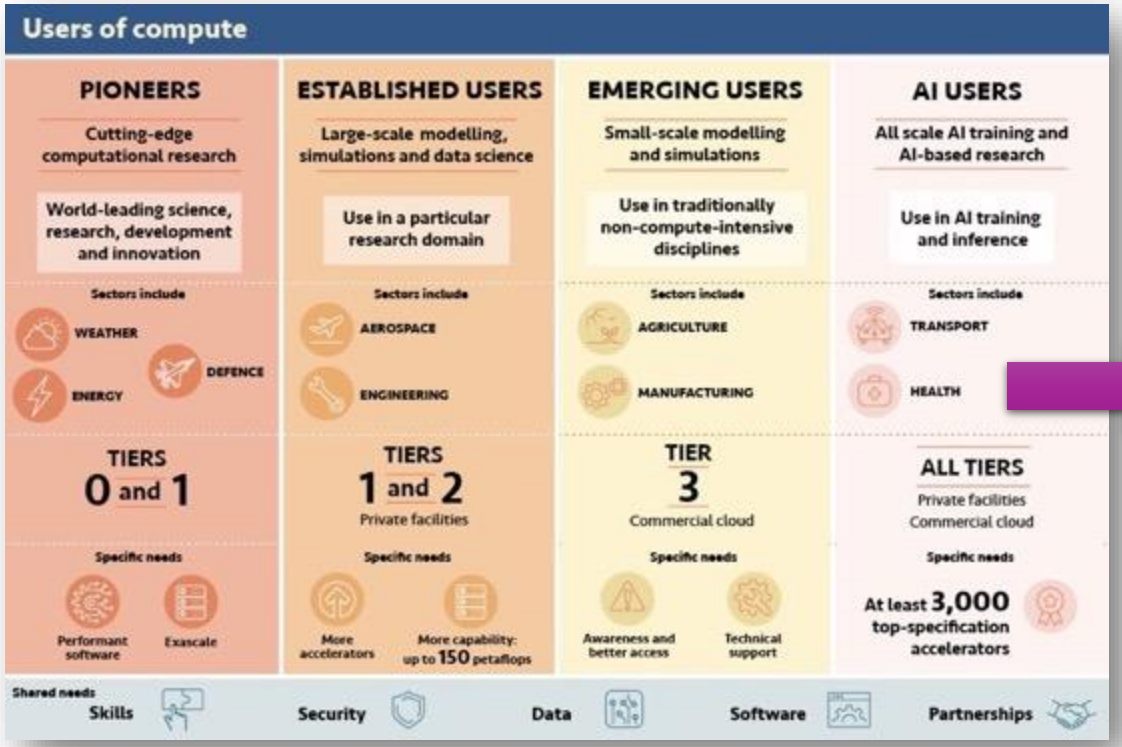
- AI is moving from computer science to computational science domains
  - AI-aided data collection and curation for scientific research
  - Learning meaningful representations of scientific data
  - AI-based generation of scientific hypotheses
  - AI-driven experimentation and simulation
- As size of data and size of models grows, capability of AI increases – leading to emergent capability
- Better understanding of how to “right-size” models to data / required capability

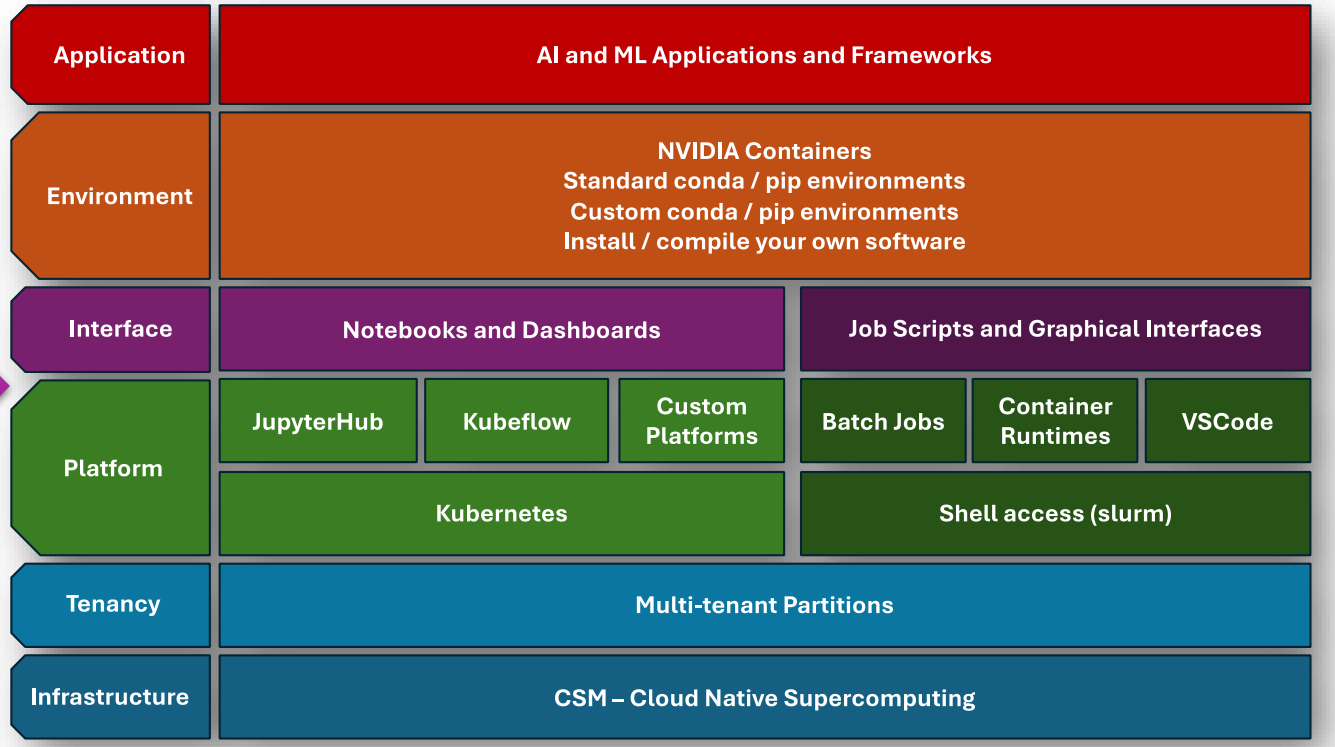


Wang, H., Fu, T., Du, Y. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).

University of BRISTOL

Addressing a key challenge—Software Stack for AI





https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-expert

Converged Cloud and HPC software stack of Isambard-AI for diverse AI and ML platforms and hybrid workflows

University of BRISTOL

Panel Q: Competition with Magnificent 7

Explainer

Why have the big seven tech companies been hit by AI boom doubts?

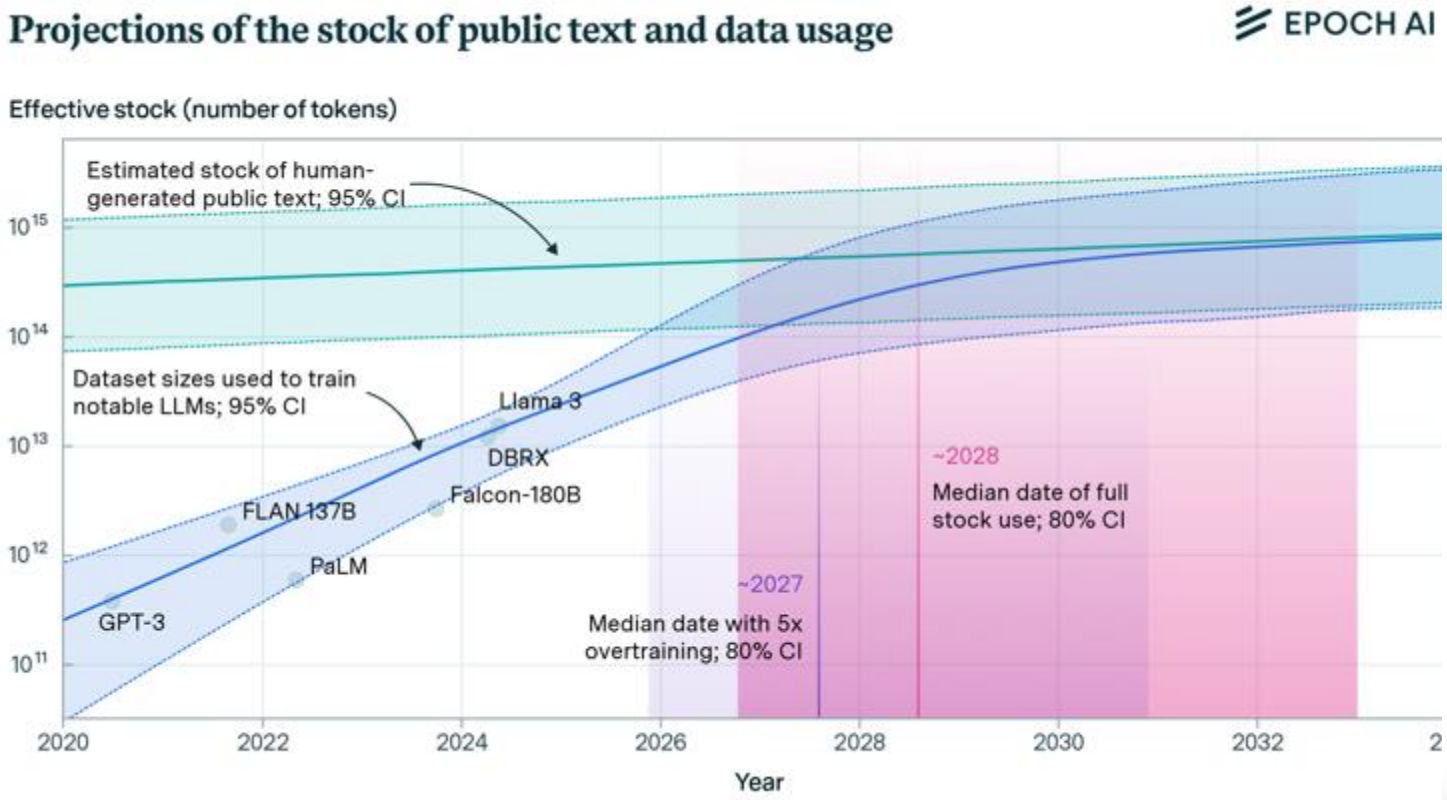
Their shares have fallen 11.8% from last month's peak but more AI breakthroughs may reassure investors

The seven companies moved into correction territory this week. Composite: Various

It has been tough week for the magnificent seven, the group of technology stocks that has played a dominant role in the US stock market, buoyed by investor excitement about breakthroughs in artificial intelligence.

University of BRISTOL

Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data



When Will The Trillions Invested In AI Pay Off? Sooner Than You Think.

BY JOSHBERSIN · PUBLISHED AUGUST 8, 2024 · UPDATED AUGUST 9, 2024

In the last few weeks there has been a lot of concern that Gen AI is a "bubble" and companies may never see the return on the $Trillion being spent on infrastructure. Let me cite four analyst's opinions.

Will Today's Massive AI Investments Pay Off?

MIT professor Daron Acemoglu estimates that over the next ten years AI will impact less than 5% of all tasks, concluding that AI will only increase US productivity by .5% and GDP growth by .9% over the next decade. As he puts it, the impact of AI is not "a law of nature."

On a similar vein, Gary Marcus, professor emeritus of psychology and neural science at New York University, believes Gen AI is soon to collapse, and the trillions spent will largely result in a loss of privacy, increase in cyber terror, and a lack of differentiation between providers. The result: a market with low profits and big losses.

Goldman Sachs Head of Equity Research Jim Covello is similarly pessimistic, arguing simply that the $1 Trillion spent on AI is focused on tech that cannot truly automate complex tasks, and that vendors' over-focus on "human-like features" will miss the boat in delivering business productivity. (He studies stocks, not the economy.)

https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

University of BRISTOL

Save Email

> Proc Natl Acad Sci U S A. 2024 May 21;121(21):e2314021121. doi: 10.1073/pnas.2314021121. Epub 2024 May 9.

Can Generative AI improve social science?

Christopher A Bail [1 2 3]

Affiliations + expand

PMID: 38722813 PMCID: PMC11127003 DOI: 10.1073/pnas.2314021121

Abstract

Generative AI that can produce realistic text, images, and other human-like out
transforming many different industries. Yet it is not yet known how such tools
social science research. I argue Generative AI has the potential to improve s

FINANCIAL TIMES

MPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI

Artificial intelligence + Add to myFT

DeepMind and BioNTech build AI lab assistants for scientific research

Artificial intelligence used to help researchers plan experiments and better predict outcomes

Science is our guide, our anchor—challenge is to continue articulating societal benefits

European Commission

Research and innovation

Home > Research by area > Industrial research and innovation > Artificial Intelligence (AI) in Science

Artificial Intelligence (AI) in Sci

Articial intelligence research, funding, policy and related publications.

Submit manuscript

ution in high-throughput proteomics and AI

nors Info & Affiliations

2024 • Vol 385, Issue 6716 • DOI: 10.1126/science.ads5749

cent capability to measure thousands of plasma proteins from a tiny blood sam-
as provided a new dimension of expansive data that can advance our understand-
, of human health. For example, the company SomaLogic has developed the means
measure more than 10,000 proteins and Thermo Fisher's Olink assays over 5400 proteins from as little as 2 µl. When these rich data are integrated with other layers of information from large patient cohorts, such as the UK Biobank's genetic, health, and

Source: reddit (two AI generated variants of horizon dawn)