HPC-AI Advisory Council UK 2026, the short version

Kolen Cheung
Research Software and Analytics Group, University of Exeter

November 10th, 2026

This is the short version of my notes from the HPC-AI Advisory Council UK conference in Leicester this October, which are linked at the end: what Rich and Gökmen told me about agents on Isambard-AI and about training, the UK's new structure of national compute, and UK-LLM.

Talking to people

It was nice catching up with Rich (Richard Gilham, Bristol Centre for Supercomputing), who works on Isambard-AI. I asked him about the agentic workloads his talk mentions. He said there are obvious signs that a lot of users are running agents, and they get things done, but also cause problems, e.g. running on login nodes. What surprised me is that it is not worse than in the pre-agentic era, just different. I would have expected AI agents, especially in unskilled hands, to cause all kinds of problems on multi-tenant systems like HPC.

BriCS's guide Using AI agents with Isambard sets out how they want agents used: run the agent on your own machine, never on a login node (that is Unacceptable Use), and copy its suggested commands into your own terminal after checking them. They also publish isambard-sc/skills.isambard.ac.uk, a plugin of agent skills, one per guide in their docs.

I met Gökmen Kilic (Durham University), who gave the DiRAC Training Academy talk. We talked about training: his experience of building notebooks and serving them, so that they are easy to deploy and to use. That has some similarities with the LFRic Atmosphere training I'm working on, with the extra challenge that our software environment is more complex, and we don't have a hosted, common platform to give users as seamless an experience. I'll see if there's anything we can borrow from there.

The new national compute

The day 2 keynote presented a new structure of public compute, which I unpacked. The main change is that tier 1 and tier 2 stop being levels of size (Figure 1). Systems are grouped by what they are for instead: the largest simulations (NNSS), AI (AIRR), and everything else on a mix of hardware (the National Compute Resources, NCRs), with one portal to apply to all of them. Isambard 3 becomes one of the NCRs, and the NCRs come in a rolling programme rather than as one-off grants to each centre.

That rolling programme matters for the people too. A sysadmin from EPCC and I talked about the instability of funding: HPC in the UK often works by writing the next proposal, and planning on the assumption that it gets a green light. Let's hope that the new structure provides a more stable foundation, both for planning the hardware and for the people, so that we can invest more in the future and collaborate more.

UK-LLM

The new structure may well give a more stable foundation, which is an improvement, but the opening keynote, by Pontus Stenetorp, shows it is not enough. UK-LLM is his project to train open source AI in the UK, and what it needs is compute for pre-training (recording):

> I estimate that the largest viable open source AI model with ethical training data requires about 20 million GPU hours, just to train it, with a little fudge factor. […] But you most likely want to train two per year […]. That number is awfully close to me saying: I want the whole of Isambard for a whole year if you want me to get to the frontier. That's not going to happen.

And the new £750 million machine for the AIRR is for inference:

> There's a problem, though: that investment is earmarked as inference only, and I cannot for the life of me understand why. […] which models are we planning to run inference on? The PRC ones? […] I had a meeting

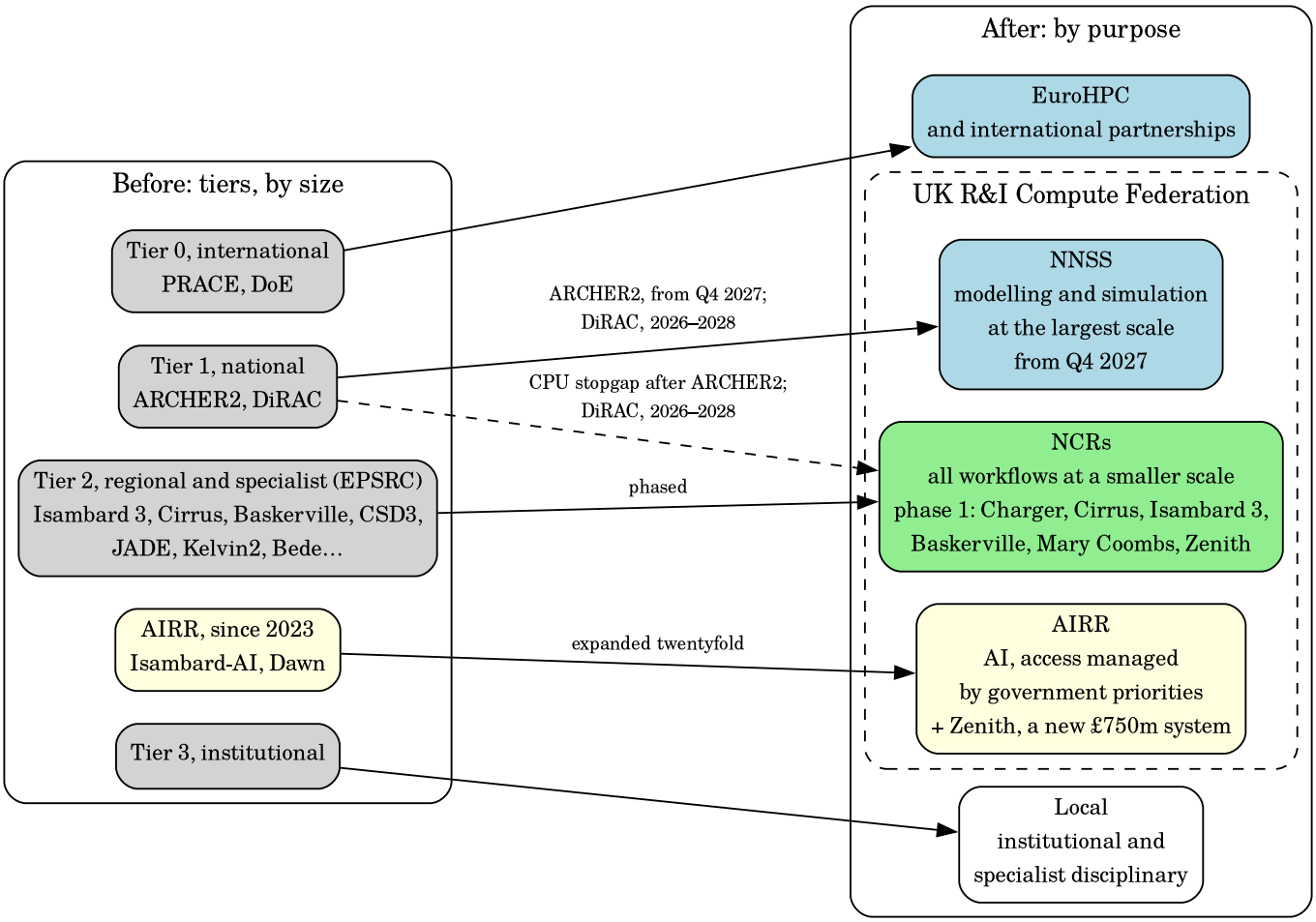


Figure 1: Public compute in the UK, from tiers sorted by size to services sorted by purpose.

with the government last week where I said: if there is no plan to build something akin to Isambard again within the next couple of years, I can move abroad, and shut down trying to do research in the UK. Isambard is not going to last forever.

Why not fine-tune someone else's model? Because "if pre-training is off the table and we leave it to the corporate world, there are scientific questions we cannot answer." I wrote about the rest of the talk in We can know, thus we must know.

Odds and ends

The sysadmin from EPCC said Research Infrastructure Engineer is a better description of his job. Of his whole cohort, only he stayed in HPC, and the rest went to quant etc., not happy, but much better paid. We share a lot of similar laments: of the friends I had who loved mathematics and physics, I'm the only one who stayed in academia, and the others went to Meta, Google, quant, AI startups, etc.

Funnily enough, Gökmen and I ran into exactly the same problem, independently, when booking hotels: we both clicked into last year's conference page, which came out on top, and booked the 7th's days, 14–15 October, a day late. Both of us went through some turmoil to fix it. Frankly, the conference's website should have a better UI/UX. They should hire an RSE to improve it.

Thanks

Thanks to CAKE, whose Placements & Visits award funded my trip, I was there for the conversations above. The talks are on YouTube, linked from my notes, but the conversations are not, and they are where I learnt the most.

Further reading

1. Summary: a map of the talks, the highlights and the people
2. The national compute: the new structure, its systems and its funding
3. Day 1, 13 October

4. Day 2, 14 October
5. UK HPC systems
6. We can know, thus we must know, on the UK-LLM keynote