



Opinions Libres

le blog d'Olivier Ezratty

Decode Quantum with Chris Ballance from Oxford Ionics

Welcome to the 82nd episode of Decode Quantum, this time, with **Chris Ballance**, the cofounder and CEO of **Oxford Ionics**. This time, I am alone, without Fanny who is seemingly busy at OVHcloud for whatever reason and couldn't be with us today. She's like a Minister nowadays.



So, Chris, to introduce you, you are the cofounder and CEO of Oxford Ionics since its inception in 2019. Your cofounder is Tom Harty who is the company's CTO. Before that, you were a senior researcher at Oxford University, where you did your PhD between 2010 and 2015. And by the way, just yesterday, you just inaugurated your new global headquarters in the Oxford Technology Park, 2,800 square meters, along with UK Science Minister Lord Patrick Vallance.

UK Science Minister Unveils Oxford Ionics Global Headquarters by Matt Swayne, The Quantum Insider, March 2025.

Chris Ballance: I'm a reformed physicist now, and now a salesman.

Olivier Ezratty: yeah, you became a salesman. At least you've got a background of physicist. So you were a senior researcher at Oxford University, where you did your PhD that you finished 10 years ago. And by the way, just yesterday, you inaugurated your new global headquarters in the Oxford Technology Park. 2,800 square meters so it seems large. And along with the UK science minister. So congratulations.

Chris Ballance: thank you, it was a great event. And we had a few hundred of our closest friends there.

Olivier Ezratty: you had a lot of friends, all your ecosystem, that looked fine. So as usual, we start with the same question. It's very predictable and deterministic. How did you land in the quantum world?

Chris Ballance: I have to go back to before I was even at university, when I was in high school. I was completely enamored with chemistry to start off with. This was post 9-11, and there was a very low threshold on how much homemade explosives you were allowed to bring into school before people started talking about expulsion. So, at that point, I decided that chemistry wasn't so great, and maybe I should go into physics. I then ended up going to Oxford University as an undergrad in physics. I was really looking forward to doing this and expecting to be around lots of people like me who liked making stuff happen.

I was slightly disappointed with what I ended up with. There were lots of very smart people, but they wanted to solve the problem set in front of them. They wanted to get the tick, get the correct answer. They didn't want to go build something and make it happen. I ended up in software, very briefly, and found that I absolutely despised corporate life. I then found quantum computing, and started a PhD in that field, which nicely married. Back in 2010, it required a real deep knowledge of physics. You really needed to understand the gory details of the theory but also which end of a soldering iron to hold, knowing how to actually build something and have that intuition. This combination was perfect in quantum computing.

Olivier Ezratty: I didn't find that in your bio online, so you had a kind of regular professional life before doing your PhD.

Chris Ballance: it was about three months of dipping my toes in, so it was very short and then deciding that this definitely wasn't what I wanted to do for my life.

Olivier Ezratty: can we say that when you did your PhD, you were an experimentalist or a theoretician or a mix of both?

Chris Ballance: any experimentalist is a mix of both! An experimentalist is someone who has a mission to go make something work and will do whatever it takes to get that working and a lot of my PhD was focused on trying to really understand where the practical fidelity limits were for quantum computing, primitives, in particular, in trapped ion systems that I worked on at the time.

Olivier Ezratty: do you remember the title of your thesis? It's only 10 years ago, so I presume you can.

Chris Ballance: all my papers around that time were high-fidelity XYZ, and I think my thesis was high-fidelity quantum operations in calcium qubits.

High-Fidelity Quantum Logic in Ca⁺, 2014.

Olivier Ezratty: so you were using calcium ions back then. It's an element that's widely used, but it's not the one you use right now. We're going to talk about that later. Back then, you did beat some records in the fidelities you could reach with ions.

Chris Ballance: back then, ions were pretty high fidelity compared to all the other quantum computing modalities out there. Fidelity is a real super metric in quantum computing. If you can find ways to increase fidelity even slightly, it allows you to run much deeper circuits if you're thinking before error correction. And it also massively reduces resource requirements with error correction. And around that time, the only gate, you know, quantum gate operations in any modality were just about floating around the threshold of quantum error correction. This is the point where, even if you had infinite qubits, adding more qubits doesn't make the errors better. So we're really interested in what the physical limits were about this. You can all swipe your pen and say, oh, perhaps if you have Majorana qubits, you don't have significant errors or whatever.

As soon as you actually start building it. any system in the lab, you suddenly realize there's a whole lot of interesting physics that emerges out of the woodwork. I ended up building out some experimental systems

where we demonstrated some incredibly low error single qubit gates. We were down at very low 10^{-5} to high 10^{-6} single qubit gate errors, showed really long coherence times, tens of seconds in these systems, and most importantly showed that we could get our two qubit gates up to around 99.9% fidelity, which was a world record back then, and at the same time showed that we really understood where each of the different sources of error were coming from by messing around with parameters and showing that the trends of performance were exactly what we expected, which broadly then means we knew exactly which levers we had to pull to improve these errors.

Olivier Ezratty: I've noticed in the history of trapped ions that they were always the best in class for qubit fidelities. Even in experimental setups, about 10 years ago, they were always the best in class for this. There's this work, all the work from David Wineland in the US and others, and the guys who led to the creation of Quantinuum and so on. But it was difficult to then scale that in production grade systems. So that's what you're going to talk about later. So how do you do that on production scale? So you've done your PhD. And then what? You then worked at Oxford.

Chris Ballance: I went all the way from being a PhD student all the way through to running my own research group and building out a team. And I was really interested in all these different aspects of scaling. And I took this in a few different directions. So I spent a bit of time working out how you could entangle different types of qubits. So we, for example, use strontium qubits and calcium qubits and work out how we could do the very first entangling gates between these different types of elements. And that's interesting scientifically, since you can do things like Bell tests and test certain non-locality models.

It's also just really cool, you know, playing around, building out these entangling gates. and really playing out in this logic. We were interested in using it for better ways of sympathetically cooling quantum logic, where you can use the best type of qubit for the best possible operation. And we had in mind the next step along this journey, which is building quantum networks. So after I'd done that, I set about building a quantum networking system. And the idea behind this is that, you know, this is the story that I told myself back at the time that I now think slightly differently.

If you can build small quantum systems, small quantum computers, and you can network them together using, say, photons, you can then build arbitrarily large systems. And it's an engineering problem, not a science problem. So what we set about doing was showing that if you took two qubits in these different devices, you could entangle them by sharing photons between them. And this had been done before a few different times by people like Chris Monroe, one of the founders of IonQ in the US. But we built an apparatus that showed for the first time that you could increase this rate by a few orders of magnitude and increase the fidelity by nearly one order of magnitude. And hence get to the point where you can use this as a primitive in a quantum computation. And that experiment still runs back in the university now. Most recently there has been doing lots of fun experiments like showing distributed atomic clocks, doing quantum operations, quantum computation spread across the network, distributed quantum computing, blind quantum computing and all of these sort of interesting things. The big challenge, and this is where all the story stitches together, is that this was fun and all, but it was very clear it wasn't going to scale up by doing more of the same. And bluntly, this is because we use lasers. And as anyone who has an experimental bet who's ever worked in a laser lab can say, lasers really suck.

Olivier Ezratty: I heard about that, and also with cold atoms, so not just for these kind of experiments.

Chris Ballance: in any atomic system, atoms are great. Since I'm a firm believer of the physicist saying that there's two types of matter in the universe. There's atoms and then there's junk. Anything that's made out of a collection of atoms is inherently imperfect and it's junk. But if you have individual atoms, you have a God-given contract that they're guaranteed identical across the universe. And that means you don't have to engineer

them, which is perfect.

Olivier Ezratty: compared to a circuit like a superconducting qubit?

Chris Ballance: exactly. Compared to anything you have to manufacture, like a superconducting chip, like a spin embedded in a silicon chip where you care a lot about the environment, about the bath. All of these things you get rid of if you have an individual atom.

Olivier Ezratty: what's funny is you move the needle elsewhere. I mean, you mentioned lasers. Lasers, they have phase noise and other problems, depending on how you use them to control your qubit. If you use it only for cooling, like you do, you remove the problem in some way.

Chris Ballance: that's exactly right. So once you have the qubit, you have a perfect qubit. This is why you make atomic clocks out of ions or atoms. Then it just becomes to the qubit is perfect and anything you do to it, it sucks. And now you have to work out... how you can minimize the amount of problems in all the stuff around it, which is the next layer of the onion of system engineering.

Olivier Ezratty: before this episode, we had **Chris Langer** from Quantinuum, you probably know him, and **Sebastian Weidt**, your colleague from Universal Quantum, also in the UK. We had, even before **David Wineland**, and even before we had the guys from France, who are creating the company called **Crystal Quantum Computing**. So you're the fifth in the row, talking about trapped ions. Our audience, if they track that, they know about some of the issues of trapped ions, but we can still recount all of that. Let's talk about the creation of the company. So you are a couple years ahead at Oxford, and then create the company. What drove you to do that six years ago?

Chris Ballance: what was on our mind ever since we did these projects? I think it was really, really high-fidelity gates. was that it was very clear you couldn't scale these kind of systems by doing more of the same. Lasers suck. You can fight lasers. You can get the phase noise down. But we were really dreaming about large scale systems, with tens of thousands of logical qubits, hundreds of thousands or millions of physical qubits. We just couldn't find a way, no matter how we did the engineering budgets, to see a way of making these systems work.

And what we didn't want to do is set off doing something that wasn't going to scale out all the way to the end game. And, you know, we essentially said that we have to find new building blocks. And we worked out in particular back then on how we could use oscillating magnetic fields, at the microwave or RF frequency rather than oscillating laser fields to do these gates. And what was interesting about this is it looked really hard. Like from a physics perspective, it looks really hard. But from an engineering perspective, it looks really easy. And the physics perspective is that a non-recurring engineering cost.

Another thing is, is once you've worked out. what kind of antenna structure you need to build onto a chip to control your qubits, you can now copy-paste that antenna structure and build a lot of them very fast with very good tolerances. We got this working in the lab. We got up to the point where we had 99.7% fidelity to our qubit gates. Lots of people call that pretty good. We call it pretty sucky, but we knew what was wrong and knew how we could get around it. And when we had that technology really working and properly understood, that's when we knew we had something worthwhile while scaling and when we started building out the company.

Olivier Ezratty: you created a company with Tom Harty?

Chris Ballance: I founded the company with Tom, who I have been working with since I started my PhD a decade ago, which is amazing, by the way. We know each other longer than the average marriage.

Olivier Ezratty: so it's a "good friends" team. When you create a company, it's very important to know the people you are working with.

Chris Ballance: exactly, and importantly, without all of the arguments we could possibly have had, before we founded the company, which is incredibly important to the founding team. and have new arguments and find new faults with the people you're working with, you want to know them and trust them.

Olivier Ezratty: usually when I'm trying to figure out how you scale with trapped ions, I'm trying to look at the pros and cons of the technology. So we know it's very high fidelities and you're going to talk about the records you recently passed last year. But then you've got the problem with scaling, the problems with the gate times, the problem with also the many-to-many connectivity. To what extent is it really in place? Can you parallelize the gates and stuff like that? So can we cover those various issues and how you want to fix them in your journey?

Chris Ballance: those are all great questions. That's basically the laundry list of all the stuff we spend most of our time talking about. First of all, you have to ask what metric do you want to try and optimize? And for us, that metric is clear. It's time to solution for a given problem. And this is what customers care about. And it sounds like it should be the obvious metric, but it's still very easy to find yourself drawn to looking for proxy metrics like overall gate speed. Because when you start looking at time to solution, you find yourself looking in two very different regimes. You find yourself looking for small scale problems on say a few hundred qubits. There, time to solution broadly matters about having low enough error rates that you don't need error correction. And you get to the point of saying you can build systems that solve the problem or build systems that don't solve the problem because they have too much error. At which point you're really focused on error rates and compute time is never going to be a limit. If you're looking at large scale systems, you're looking at your error correction cycle time. And then you see quite a few different variables that matter. So you care about the physical gate speed, slower in ions than other systems. So you care about connectivity, much better in ions. You can do non-local codes. You care about error rates, much, much better in ions, which massively reduces the quantum error correction overhead. You care about the parallelizability of the gates. And in our architecture, we can fully parallelize all of the operations. And then when you add all of these different aspects up together, you find out.

Olivier Ezratty: how about readout? It's not a given everywhere. Sometimes, readout is 10 times longer than the gates.

Chris Ballance: exactly! You need efficient parallel readouts. That's a given. And once you have. But when you have all of these things put together and you start doing your resource estimates, you then find that at least, say, looking at, say, superconducting qubits or silicon spins, you end up at roughly the same time with a few factors of two in either direction, depending on how you cast things. And the way I read that and the way we read this back five years ago when we were probably starting to build a company is that actually it doesn't matter so much which of these technologies you pick. It's a matter how efficiently you can scale them and how efficiently you can build engineering-scaled teams to solve the problems and to separate the problems. Rather than ending up with a tangled plate of spaghetti, we are solving fundamental... physics problems left right and center and you can't ever separate off things into black boxes and set a team of people off to go solve that problem while knowing it's definitely solvable within that that boundary that set of resources there's a problem with land engineering programs is that if suddenly you know joe's team over there has to start changing its interface which then changes alice's team over there suddenly everything falls apart everyone blames each other and you can't get anything working you have to find ways of tightly scoping what's possible giving a little bit of buffer on each of the different module specifications and then getting on with actually.

Olivier Ezratty: what you describe is the problem for all qubit modalities. It's a matter of global optimization

to some extent. You have many different figures of merit and many parameters and you need to create a model of everything including algorithms and quantum error correction and whatever and find out which parameters you tune. It's not specific to ions. It's a very global problem.

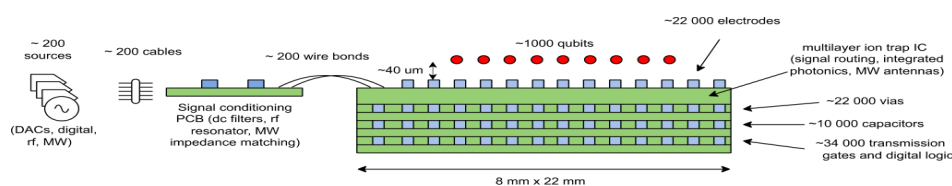
Chris Ballance: what we often say internally is there's a lot of architectures out there that look really elegant to physicists. They draw them on the whiteboard and you have your perfect array of qubits. You draw everything out and it looks beautiful, really symmetric, beautiful, sensible. We don't care about building systems that look elegant to physicists. We want systems that look elegant to system engineers. And if you can't find a way of finding your system so they are elegant to system engineers, you then end up getting stuck in engineering hell.

Olivier Ezratty: to give credit to other companies, they all have engineers and they all have people doing the integration of all the aspects, not only physicists, but probably younger startups have more physicists than engineers, but more mature companies.

Chris Ballance: this also matters about how you think about the architecture, since you want to be able to separate off the physics packages from the chip design packages. And then our scheme, say, where we use electronics to control the qubits, these electronic signals buried in the chip, then the physics, the quantum stuff is just around designing a way, designing your gates, which you can test on very small. chips. It's then about designing antenna structures into the chips that generate the appropriate E-fields and B-fields. And then it's about modeling out the crosstalk and the interactions between different zones. And the magic here, which is where all the system engineering gets much, much easier for us, is that our chips are then purely classical. These micron-scale features and they have macroscopic currents. This means a couple of awesome things. It means you can build them pretty damn easily without too much sweat. You can model them using industry standard tools, which then means you can hire out relatively industry standard teams.

Olivier Ezratty: you mean EDAs, the software tools you use to design your circuit are regular. You don't need specific EDAs like for superconducting qubits and stuff like that.

Chris Ballance: that's exactly right. Then you can hire people, you can build out industry standard tools, which then means when you're talking with the foundries, they know exactly what you need. When you're hiring out scaling teams that deal with the different aspects they all need. So there's still tremendous numbers of hard engineering problems. But they're hard engineering problems that you can attack using ways that everyone in the industry knows how to attack them. So then it becomes something you can coordinate a lot better as an engineering program, rather than having to reinvent a whole industry from scratch and from imagination.



How to wire a 1000-qubit trapped ion quantum computer by M. Malinowski et al, Oxford Ionics, PRX Quantum, May-October 2023 (21 pages).

Scalable, high-fidelity all-electronic control of trapped-ion qubits by C. M. Löschnauer, C. J. Ballance, C. Matthiesen, M. Malinowski, T. P. Harty et al, arXiv, July 2024 (12 pages).

High-fidelity heralded quantum state preparation and measurement by A. S. Sotirova, M. Malinowski, C. J. Ballance et al, Oxford Ionics, arXiv, September 2024 (17 pages).

Olivier Ezratty: let's come back on one very simple question. So you use barium ions. There are some other options. People are using calcium, ytterbium sometimes. What are the driving choices for this ion? If I'm not

wrong, you use only single species system. Some companies are using dual species for various reasons to cool some of the atoms and use the other for qubits. So what's rational between your choices over there.

Chris Ballance: they're all engineering choices, not physics choices. So if you look at the engineering system, barium, by far the best qubit. It's nice and light, so you can drive it very hard using electric fields. You have to use UV wavelengths, which absolutely suck. For state preparation and measurements. If you look at calcium, it's very nice. The lasers are very convenient, even more convenient than barium, but it's very difficult to build light-carrying structures on a chip, so integrated optics on a chip to deliver the state preparation and measurement light. And barium then broadly misses all the metrics that physicists care about. It's pretty much the worst choice in all of these, by a factor of two or so. But it turns out you can then use certain industry standard materials in designing your chips, which are already in existing foundries. So what this means is that it's just in the right sweet spot, where you don't have to put large amounts of effort in your chip design to make these things work. You have to put more effort into the physics design, but again, that physics design is a non-recurring engineering cost. It's a problem you solve once. So this is really an engineering-led decision of how do we minimize the global cost of the system and making sure the stuff we have to engineer a lot of is easy, robust.

And it's not just about cost and cheap using standard supply chains, since doing new things in, say, fab supply chains and foundry supply chains. It's much, much harder than doing new things in a physics model once you have sufficient understanding of the moving parts.

Olivier Ezratty: I remember that about three or four years ago, IonQ chose to switch from ytterbium to barium. One of the rational was you can drive them with lasers that are compatible with a telecom wavelengths at around 1,500 nm probably, something like this. So was that rational as well for you, or is it less of a problem for you to manage interconnect and using telecom wavelengths.

Chris Ballance: it's part of using more standard materials, which have more standard supply chains. It's very similar engineering considerations. How do you minimize the overall total pain in your supply chain? A lot of which comes from using stuff that's not standard. Ideally, you want to be following a trillion dollar industry like a semiconductor industry and just doing what they do, using the tools they do. Since you can build new technology for an existing market or use... use existing stuff that works for a new market. If you're trying to build completely new technology, completely new supply chains in a completely new market, then you're in for a really hard time. You don't want to be doing that.

Olivier Ezratty: I presume that you cool ions and control them at the beginning with a laser. Then you put them on top of a QCCD circuit. These QCCD circuit have been in place for a long time, about 20 years. So you've got a 2D layout, if I understand well, on your circuit. So what's special on the QCCD is the way you design it in multilayer. You use only CMOS manufacturing techniques. What other special things do we have in that circuit to control your ions? There's this microwave thing that is important, I guess.

Chris Ballance: that's right. If we look at the processes from the inside of the in and out, you have your arrays of qubits, ions, which you can load into the device easily. And because you have very deep potentials and you're in a good vacuum, they just stay there for a long time. You've had, want to move these qubits around, so confine them in potential wells, reorder them. That's done with top metal, so electrodes on the top of the chip that generate electric fields. And then when you put the qubits in certain positions, you want to control them.

And that's then done with antenna structures, which have currents passing through them, which then are layered below that in the chip. And then further on down the stack, you eventually have integrated optics, which delivers things. A lot of the magic is when you start finding the right solutions to these problems, everything starts falling out nicely. So you can start finding very clever ways to reuse, most of the current flowing in the

chip in a global way, along with your electric fields, in order to remove a lot of the control knobs you might otherwise have to do, to essentially turn the problem into one of duplication to build additional cells. And then, for example, rather than having to turn light on and off in different parts of the chip to reset the qubits in different regimes, you can start globalizing these contingencies. And that's what we're going to do. Control feels better by, again, having these unit cells. And what we've worked on how to do is find solutions that allow us to tile out very standardized repeating elements from the chip to make larger and larger devices. So our 256-qubit chips look very similar to our 32-qubit chips, which look very similar to the eight-qubit chips we published data on last year.

Olivier Ezratty: let me rephrase what we described. I like to be like an engineer there. So you've got a circuit which is conveying which kind of signals. So you've got microwaves. I don't know how many gigs it is, but so, okay, microwaves. So we've got microwave guides in the circuit at some layer. You have light, you said. So we've got some laser light coming in waveguides there as well or in the circuit. You mix lights and microwaves. And do you have DC current as well or some other kinds of current?

Chris Ballance: no, it's just voltages to generate electric fields, curves generates microwave antennas, and then light to do qubit state preparation and measurement. So there's a really important separation here, because with qubits, quantum computers, you do two different types of operations. You do dissipative operations, like reset, where you want to chuck away information, like measurement, where you want to collapse the quantum state. Those are robust, because you typically don't care if you apply twice as much reset to your qubit, they're still equally reset. So you have very loose tolerances, like 20 or 30% tolerances on those fields. Then you have your coherent control, which drives your quantum operations, and their tiny imperfections in those control signals lead to imperfections in your gate, which lead to errors that accumulate rapidly.

Olivier Ezratty: can we separate here what is a global control and what is local control in the circuit?

Chris Ballance: it turns out that almost all of the state preparation and measurement steps you can do with global optical fields, which you can deliver efficiently. with then localized electric fields only. Likewise, with a lot of the coherent control, you can do tricks where you turn all of the control issues, into electric field control, so quasi-static electric fields, rather than magnetic antennas. So in fact, speaking concretely on some data we published last year, so I can talk about more easily, we have a 10-qubit class device, where we show that by displacing the qubits radially, slightly, above this magnetic antenna, we can take this array of qubits and turn on and off the individual qubit controls, where we show crosstalk errors down at the 10^{-6} level and single qubit gate errors down at the 10^{-5} error per gate level. And the only thing that's unique, that's addressable, that's local in this device, is changing a few hundred millivolts and some DC patches of antennas.

Likewise, we showed in that paper, we can do two qubit gates, and then by changing, we can do two qubit gates, again, the electric fields on these electrodes. You can change the frequency, the trap strength of these confinements, these potentials, and use this to turn on and off two qubit gates or vary the angle of two qubit gates in different places. So what this means is that the only control you need locally is changing a few electrodes by a few hundred millivolts. And what's cool about that is, first of all, these are large voltages. These are macroscopic voltages. This is very easy to generate intolerance. And then the trickier signals, the microwaves, are global, which means you don't have to worry about cross-talk, or rather you can model this out in a periodic lattice-based approach where you understand what's going on and can really control it well.

This means you have these separable problems that are nice. And the other really interesting thing about this architecture, when you take your systems engineering hat on, is that modeling this, modeling these electric fields, that's a really standard CMOS MEMS type problem. There are just incredibly good commercial tools. They can do this incredibly well. And then you can easily Monte Carlo against this to analyze and tolerance

what's going on. And you're really well within industry standard models. So you have a design problem. You need to work out what question you want to ask of your design. But you can definitely build these designs well with good tolerances. And this means that as you start.

This means that as you start building larger and larger systems, you can make pretty good design contracts where you can tolerance all the different things going on. And really build large systems without having to add lots of unique knobs, which then becomes important for one of the things you mentioned early on. That is a key word that's very important to us, which is parallelism. How efficiently can you do stuff in large devices without having to have a tyranny of knobs? Since it lets you want to build a thousand qubit device, you can't have five thousand knobs you have to tune up to make that device work. If you do that, you're in for a very sorry state of affairs.

That means you can never build a hundred thousand qubit device by doing more of the same. In lots of our devices, almost all these parameters are baked in at the design. So you can tolerance devices and essentially you can swap out devices and they all work the same. And this is just a couple of global control parameters you need to calibrate on a device by device basis. And that's ultimately why we're so excited about this architecture, since the tolerancing problems are stuff you have to worry about at the design time, rather than on a device by device metric. And we've now shown this across hundreds of devices that we really can to swap them in, swap them out and see identical performance between the devices to the point where the science team don't know if they're running on a different device or not, or a different system or not, because, you know, the technician team can swap these things around. You have to go right and dig into a database and work out the serial number of the device you're using to work out if it's the same one as you're using last month.

Olivier Ezratty: can we come back to the ions themselves? What is the ion manifold? So what's state zero and state one?

Chris Ballance: we use stable states in here where you use two different hyperfines. states. So spin down is a lower hyperfine state, and spin up is an upper hyperfine state. So this turns out... And separated by a microwave energy level. That's it. It is separated by a low-frequency microwave level.

Olivier Ezratty: what's the control frequency by the way?

Chris Ballance: about 1 gigahertz.

Olivier Ezratty: that's typical for hyperfine transitions.

Chris Ballance: they can be a magnetic field across the system, such that you can now tune to a first order so as a first order zero, a first order null, so a quadratic part of the magnetic frequency spectrum, which then makes you insensitive to any small magnetic field fluctuations. So like this, you're in what's called a clock qubit, which is the same kind of qubit used in atomic clocks where you have exceptional coherence times. So back in the university, we really pushed hard to measure a proper coherence time on this. We ended up measuring in a system where we had a 10 minute, coherence time. And interestingly, our coherence time wasn't limited still by the qubit at that point. It was limited by the atomic clock, atomic reference we are comparing the qubit to. So when you start tuning up these qubits, they really are perfect. And again, all that goes wrong is the fields you use to control them. So you have to now tolerance these classical fields well.

Olivier Ezratty: the second \$100 question is about scale. So usually when I try to find out how you scale, I separate scale in, so how large is the processor, and scale out, at which point do you need to connect multiple processors and how you do that? So can you separate those two things?

Chris Ballance: first of all, I'm one of the world experts on networking together quantum processes. I think the

results from the university from 2018, 2019 still hold the world record for a quantum network. It means a lot when I say with confidence that networking isn't on our roadmap for a very long time. Because building heterogeneous technologies is really hard. It's far better to be really good at doing one thing, homogeneous technologies, than to try and have to be good at quite a few different things simultaneously. So because of the way we build out these devices, we can build out into very large dies without having to change anything. So we can get to hundreds of thousands, if not millions of qubits by carrying on with the same processes we're already using, just building larger and larger devices.

Olivier Ezratty: what do you mean with large? It's a wafer scale chip, like with Cerebras or something else?

Chris Ballance: large means something like a quarter of a wafer, an eighth of a wafer. And in practice, we think a lot about how you manage different fab steps and how you manage different processes to get out to suitable yield. But essentially, it means building out a device that's a good fraction of a wafer to get out to a million-fold qubits or so.

Olivier Ezratty: would you be dependent on the viability, because it's a key concern for superconducting qubits when they do larger chips. So is that less a problem for you on CMOS technology.

Chris Ballance: right, and that's what's wonderful. The answer is no, because we're using, essentially, if you think now back down to the tolerance stack, we care about electric fields at the ions. The ions are something like 40 microns above the surface of the chip, which means that as long as your metals in the right place are better than about a micron, you're happy. So this means on the critical steps, you're using something like sub-micron resolution, but not much sub-micron. And this is so training-edge in semiconductor fab, you can get incredibly high yields with it. And you can also do really good optical inspection of these devices in situ in the fab. So this means devices fail out fast in fab. You don't have to go all the way through to doing quantum tests or final inspection tests to do this. And the yield is pretty exceptional on these devices, because anyone who can build stuff with slightly finer resolution already has much better process control than we need. Thank you. So staying that few steps behind leading edge, means you're using the technology that everyone finds boring. And a firm factor, you know, my engineering belief is that you really want to be using technologies that the engineers find boring. You don't want to be using technologies that engineers find exciting. And this has stood me very well.

Olivier Ezratty: what's important is that in your CMOS circuits, you don't need EUV or stuff like that. You probably need a regular resolution, a deep UV and stuff that existed 20 years ago. You mostly need to do some 28 nano.

Chris Ballance: exactly. If you now think in the standard semiconductor chain, a lot of the late steps for the top metal are exactly the same kind of steps you need. And people already worry a lot about yield for that because you do a lot of expensive value adding steps beforehand. So if you fail out chips of those last sets of steps, it gets very expensive. So there's already people worrying a lot about yield for these kinds of steps out there in the world, which means we don't have to and ultimately this is because you know in a superconducting device you're doing stuff that no one has cared about optimizing yet at all and you're having to try and build all of that yourself and be incredibly clean with interfaces and then our devices we say we're 40 microns away from all of that so any of the kind of atomic level junk on the surface we don't care about we just care about these macroscopic features which is exactly the same as any classical logic or any classical MEMS cares about and this means we have to reinvent the technology.

Olivier Ezratty: if I was comparing your technology and many other qubit modalities, you have a scale-in ploy, with no scale-out plans at this point and you are confident that you can scale to enough physical qubit to build enough logical qubits to build reasonable usable applications. So on your maximum size single chip, how

many logical qubits could you accommodate? Or there's another way to count that maybe in operations like a teraquops. So did you account for that, including all the overhead of error correction and magic state distillation and so on.

Chris Ballance: it depends exactly what kind of error rates and what kind of modeling you're doing on this and what assumptions you make. But we're into the tens of thousands of logical qubits with very low error rates, you know, well, well, sub 10^{-10} on single wafer devices. We also have plans to then build out onto multi-wafer devices for this. Broadly in our technology, that's not so exciting. We know how to do it. And the question is when, not if. But importantly, we don't have to do this until we're many cycles down the product iteration level.

Olivier Ezratty: what do you mean with multi-wafer? So you mean that you're going to? Use wafer next to the other and use ion shuttling between wafer? like what Universal Quantum is doing or something else.

Chris Ballance: broadly, this just becomes a tolerance problem. And ultimately, what that becomes is a gluing problem. How do you glue wafers in a reasonable alignment. It's very easy to tolerance this out pretty tightly.

Olivier Ezratty: and so, viewed from the ion perspective, it looks like a giant wafer. But the manufacturing is built on multiple wafers. It's like Lego.

Chris Ballance: that's exactly right. So it becomes a tolerancing problem. And then you choose where you want to put the pane. And what's nice is you can see this trending out nicely in the scale out. So as you scale out and build more and more devices and larger and larger devices, you can just watch how your yields are trending. Ultimately, what's kind of cool is that semiconductor wafers are incredibly cheap. So if you look at the cost of supply of a system and the cost of the wafer in that, you can actually deal with very low yields as long as you can select them out efficiently before you do value adding steps on the devices.

So what this means is you have a very long time before you see yields starting to become something you worry about to the point where, starts to actually change your economic model at all, which then means you're in the nice point where you're just watching these things trend and, you know, recalculating on the spreadsheets when you have to do the slightly more painful things that have a known amount of pain. And we see, in fact, this across lots of different aspects of the tech stack. There's enough robustness, there's enough margin, there's enough overhead, that it's not everything has to work right at the beating edge. The difference between how well it can work and how well it needs to work is large enough. You have engineering margin, which is why the whole work so well in the first place.

Olivier Ezratty: you may know, but I'm also one of the co-founders of the **Quantum Energy Initiative**. It's a group of researchers trying to look at the energetics of those systems. How do you look at this problem and how about the electronics of your system? So what's the cost of the electronics driving the circuit? There's some multiplexing involved there. So that's some interesting tricks. The other one is cryogenic. So do you need to cool part of the system at 4K? So two things: electronics and cryogenics.

Chris Ballance: in terms of electronics, the systems are ridiculously boring on this in a really good way, boring for an engineer's way. Since we have these repeating uniform cells, a lot of the control signals we need to repeat. So then you can be pretty clever about how you connect together these signals to multiplex them together, at which point you need very few signal sources and just some very simple cryoelectronics, some multiplexes to make this work. We published some analysis of this a couple of years ago now. And then you can end up just using a few local control signals, which again can be multiplexed and held on sample and all capacitors. So what this means is that the number of control signals you need ends up looking like a few hundred plus something logarithmic in the number of qubits.

Olivier Ezratty: do you have a kind of estimate of the power need, kind of fully burden power need for a single qubit? Physically qubit. Is it easy to compute? There's some mutualization there, but is there a way to compute that easily.

Chris Ballance: yes, you can compute that. It's actually a slightly more complex function. There's an offset and then a very low scaling per qubit. So if we look at the total power consumption, of, say, concretely a 10,000 logical qubit device, you end up order 100 watts of power dissipation on the QPU, on the chip.

Olivier Ezratty: You mean the chip itself? I mean the electronics driving the chip.

Chris Ballance: right, exactly. So that's in the quantum part of the system, which then impacts some of the cooling questions we can go into in a second. Then you look at the electronics around it and that only ends up being a kilowatt or two. This is very simple electronics and you just need a few thousand channels to make the whole shebang work. So then when you put the whole system together, the power becomes the power dissipated on the chip, plus the cooling power, and everything else becomes small. And then the question is, what temperature do you need to work at? And this is, again, where engineering pragmatism comes into play.

So the ion physics doesn't care what temperature the chip's at. You can run your ion trap at 100 degrees above room temperature, and it works just fine. What you then care about is vacuum. The ion trap doesn't care about the temperature. The ion cares about seeing good vacuum to avoid collisions, which can corrupt data. So then the question is, how do you get good vacuum? And you can do this one of two ways. You can be incredibly precise about your packaging, about materials choice, or you can just cool it down to something like 10 Kelvin, at which point all the junk freezes out, and you can leave a ham sandwich in there and you'll still get perfect vacuum. So we cool these things down to kind of like 10 Kelvin. And what's also about that is that when you look at what you need to cool these, you know, 100,000, you know, qubit chips down to 10 Kelvin, you end up with, you know, a few tens of kilowatts of cooling power, of cooling electricity going into this to generate the cooling power you need on the device. That ends up being very similar to the machines you use to cool an MRI scanner. That is, you're in the point where you can just use industrial off-the-shelf devices that are produced in bulk, and you can order on a four-week lead time.

Olivier Ezratty: you mentioned 100 watts of cooling power, but what was the temperature? It was the temperature of the circuit, so it was 10K or something higher. So the core bits around the ion trap go down at about 10 Kelvin. Okay, so you need 100 W of cooling power at 10K, which I guess is reasonable because you've got good efficiencies at this temperature. So you need a couple of kilowatts.

Chris Ballance: exactly. It looks relatively easy. And importantly, there's some very nice maps from lots of different manufacturers that map out exactly where the plant is and you can move this up off the shelf. So even if you end up going 10 times higher, you're still talking about something you can plug into a three-phase plug in the wall and turn on that you can kind of order, probably not on your personal credit card, but on your corporate credit card and turn up within a very reasonable amount of time. So again, none of these individual aspects are hard. What's hard, and I always like to say this to the team, is that the whole thing about quantum computing isn't the quantum part, it's the computing stuff.

It's how you build up architectures where you don't have to reinvent everything every time you scale slightly or hit an unexpected bump in the road. And that's all about happiness.

Olivier Ezratty: there's some costs that we didn't mention because it may be significant in some architectures is the cost of error correction. I mean, the classical part, error syndrome detection and stuff like that. But since your gates are kind of slow compared to a solid-state qubit, probably there's less constrained on real-time error correction. Maybe do you work with your colleagues from Riverlane in the UK, I don't know, but they do that

very well.

Chris Ballance: that's a really important question as well. So first of all, you know, to take a step back and just talk philosophically for a second. I always find it interesting the language people use around error correction, which generally assumes you can't get your physical errors lower. Because ultimately what we want is to have really low errors on our qubits. And you have two levers to pull this. One is how good is your physical hardware? And then otherwise, how much can you fix in software with quantum error correction? And you end up getting massive yields, at least until you get all your gates comfortably below 10^{-4} per gate by building out better hardware.

And what this means when you start off doing your error correction with 10 to the minus four, two qubit gates and really long clearance times, and also good connectivity is you can use very efficient codes. You don't get stuck in nearest neighbor connectivity only. And also the amount of error. So that's the kind of error correction you need. the kind of the code distance you need is way smaller. So this suddenly means rather than saying, I have now a 10,000 to one encoding overhead, which is another way of saying it, is I have a 10,000 bit string that I have to work out very quickly what physical error I had and what operations I need to apply. When you're doing this with like a 30 bit string, the problem goes from being really hard to pretty naughty.

Olivier Ezratty: I've seen that there was a paper done by the guys at DeepMind, which was about surface code syndrome detection using the **AlphaQubit** software. And it doesn't scale beyond surface code of distance 11. So it boils down to what you explained. You need a low distance surface code to make it easy to decode. And so that's your plan?

Chris Ballance: once we have that, plus the really long clearance times we have, it then means you can be a lot more intelligent about how you schedule all of this, which means you don't have to do any of our correctional work and large scale systems on ASICs. We can do it all. on a relatively small quantity of FPGAs. And what's the other really cool thing I should point out at this point, which we've only recently appreciated just how awesome it is, is that the way our control fabric works on our processes, it doesn't really change. If you run different error correction algorithms, you end up with factors of two differences in speed, depending on exactly how you arrange tiny different aspects of the device design. So this means you can optimize it for any one given code, but you can also pretty much in real time, if you want to, start using different codes. So you're not then tied into doing it one way. You really have the flexibility to do software updates in your system and change what error correction code you're using, or even change it mid-circuit if you're trying to solve different parts of problems where different error correction codes are useful at different points in time.

So all those different bits together means that the hard job of really finding and analyzing exactly how you space out all these different things is something, you can tweak and iterate rather than having to pick one and then spend two years building. And that's incredibly valuable.

Olivier Ezratty: I want to come back on the energetic thing for one reason. Something that nearly nobody discusses in this world is the economics of QPUs. My kind of threshold is what's going to be the cost, I mean, the hardware cost or the usage cost and the energy cost of 4,000 logical qubits. That's my kind of threshold because I know we can do chemical simulations with that and break RSA, which doesn't matter, but it's the same kind of level of resources. So what would be your threshold? It looks like you could need less than a couple hundred kilowatts to do that. How about the price of the machine? Because I've heard, not about you, but about other companies, of prices which are in billions, which doesn't make any sense. If you want to do any computing, a billion dollar, machine is nonsense. It looks like with the kind of technology you're developing, hopefully you're not the only one there, but you could produce very affordable machines. What's your view on the economics of those systems.

Chris Ballance: so my view is that unit economics is something we think about a lot in large-scale systems, and we use as a big driver of how we develop technologies. We're very happy paying large, non-recurring engineering costs to work out how to do stuff. But yeah, the unit economics has to be good. We're really very excited about unit economics. You know, to give an idea in round numbers, the kind of costs to assemble post-NRE of, you know, a 10,000 logical qubit system, it's like \$10 million, sub \$10 million.

Olivier Ezratty: it is in the low range of any predictions so far.

Chris Ballance: right, exactly and ultimately, this is how it should be. You know, the parts we're using are relatively simple. The chips are cheap. They don't add anything to the unit economics. And then you're paying for relatively standard electronics around it, relatively standard packaging. optical setups. And there's no real one big long tentpole that gets expensive. So there's lots of moving parts you have to solve. But when you get that solution working, the cost of reproducing is low. And this is actually really important and another aspect of speed that you mentioned earlier on. And I find it interesting that lots of people don't talk about some of the embarrassingly parallel issues we have in quantum computing, where you're typically not just solving one very long compute sequence, which means you have to wait for your computer to end. You're often running 10s or 100s or 1000s of problems in parallel that are all somewhat related to get to the answer you want. So at that point, you can start running 100 different systems in parallel. And then what matters more than headline speed is like solutions per dollar hour, at which point having systems where the cost of building it is much lower means, you know, the equivalent \$1 billion compute has much, much higher compute power and compute bandwidth. just because you can buy a lot more machines for your billion dollars. And when you start seeing together all these economics of power consumption, of speed, of scale, of size, in my mind, what really matters is how efficiently can you reproduce machines? You know, how fast can you scale up production? If you have one that's amazing, how long does it take you to get to 100 that are amazing deployed out there in the market, or 1,000 or 10,000? And all of these different aspects depend so much on the unit economics and the supply chains. And I think a lot of the industry is just starting to wake up to this.

Olivier Ezratty: it's funny because **I've got a slide on that.** I show the power, not the energy, the power consumption for existing HPCs. I mean, the largest one in the U.S. is El Capitan, is about 40 megawatts. And then I plot estimates for QPUs, for 4,000 logical qubits. There's two orders of magnitude difference between the designs. Based on your simulation, you're in the low end of that range, which enables you to parallelize on multiple QPUs the same circuit chart, like for QPE, and maybe not by 100 machines, but maybe 10 or 20, that would make sense.

Chris Ballance: to get in line in the price of a top-range HPC, that would make a lot of sense.

Olivier Ezratty: so you mentioned manufacturing. Can we talk a little bit about your partnership with Infineon? Absolutely. It's a key partner for you, I understand. I know a couple of people there. I met them in Germany. So it's interesting because Infineon is becoming one of the leading companies for manufacturing chips for trap clients. Not just for you. They serve many other vendors. So what's special with those guys? Absolutely.

Chris Ballance: the way we always think about manufacturing is we don't want to be beholden to any one manufacturer. We want to have a group of different manufacturers we can work with to partner this. But we also want people who deeply understand the technology and have good business reasons. do it. And what we see in Infineon is, you know, a strong partner who has the ability to manufacture what we need and also has and is working with quite a few different customers in the space, which is awesome for us since it means the cost of their developments are now spread over multiple customers. And they have now multiple different parts in the supply chain, so they can do this. One of the big challenges with any sort of marketplace setup like this is precisely getting to the point where there's a few different vendors and a few different suppliers you can all

cross-fertilize. So you end up in a stable regime. And I think we're now starting to see this with ion-trapped quantum processor manufacturing. That's very interesting.

Olivier Ezratty: how about your customers? You started to explain that you're going to ship to NQCC a system named Quartet. What's going to be the specs of that system? So how many qubits? what's going to be the use case, probably in the NISQ regime at the beginning. Then you have this deal with the CyberAgentur in Germany, which is kind of weird because you talk about portable computer computing. So what are they going to do with that? What is the meaning of portable for you.

Chris Ballance: the portable aspect is really interesting since from a national security perspective, if you look at high-performance compute, typically they end up being relatively localized, but you have quite a few. So people's national security computers tend to be physically distributed over lots of different sites. And the CyberAgentur are quite forward-thinking and they're interested in what happens as they're starting to adopt quantum computing for some of their use cases, but where there's only a few of these systems out there. And they broadly see that a lot of quantum computers right now are kind of built into a building rather than wheeled in on a crate. And hence that it takes one bad actor, a Molotov cocktail or a pair of secateurs to take out a lot of national security infrastructure. So what they're really interested in is not quantum computers that can sit on the back of a Jeep and be driven around the field while working, but a system you can install like a server rack. So you have someone put it on a pallet jack, install it, put it on the floor, plug it in and work.

And this is where our electronic control is great for this, since the number of moving paths is pretty small. The systems are pretty robust and pretty reproducible. And we're building a system that turns out to be about a single cubic meter in size. And it draws less power than an electric kettle while holding 64 really high performance cubits. And then for the UK's National Quantum Computing Centre, we're delivering a system to them in the next few days, actually. As you mentioned in your introduction, we opened our, had the grand opening of our office with the system in kind of big display in the foyer. And that's now being packed up as we speak to go over to the National Quantum Computing Centre. And this system is designed to be upgraded all the way through to 256 qubits, you know, over the coming few years. So we have those first 256 qubit processors in design at the moment.

Olivier Ezratty: I presume you've got a roadmap for scaling up to those zillions of logical qubits. So most of the time people think that it's easy to move from 100 to a couple of million physical qubits. There are a lot of steps and a lot of difficulties you encounter. And each time you add one order of magnitude, more or less. So do you plan to scale up very fast or is it going to be a 10 year roadmap.

Chris Ballance: that's a great question. I think the answer is both. Another way of saying this is we have a lot of devices, a lot of our test bandwidth, kind of all in-house R&D systems. And a lot of our bandwidth goes to building out systems, you know, along test vectors. So we put a tremendous number of... qubits on and look for one particular aspect. Because ultimately, the way you build really large devices is not by taking very small steps. It's by taking very large steps in a few different directions to range find and sort of turn over the different stones and see what's underneath them. And then once you know, you can postulate a lot better. So a lot of the work we're doing now is laying the path to scale all the way to those few hundred thousand physical qubit devices.

Olivier Ezratty: so coming back to the previous answer you gave about this portable QPU, what's interesting is in what you describe is it matches exactly the requirements of the data center for the cloud. Having a rack mountable system, means it's going to be, easy to install in data center. I presume it works on regular data center temperature, about 20°C. So is that the case? There's a lot of constraints when you install some stuff in the data center because there's some vibrations over there. There are many other dust, time. So do you have a hardening kind of room for checking that your system is going to work in those situations?

Chris Ballance: this is something we worry about a lot. So all the systems you produce right now, the system, for example, for the National Quantum Computing Center, the upgradable system, or the system for the German Cybersecurity Agency, both of these designed to fit in standard data center footprints, which includes not just the physical, you know, can it fit? Does it fit in a standard size rack? Does it have appropriate full loadings? Is it appropriately temperature sensitive? Is it appropriately vibration sensitive? But also to the power points and the data points that you mentioned, can you easily and efficiently integrate it? And luckily with our technology, that's not so hard, since most of the stuff outside of the quantum processor is relatively standard off the shelf kit. So you have the pain of integration, as you always do, but it's integration, it's not having to build truly new things.

Olivier Ezratty: what I observe in this world of quantum computing startups is many companies have a tendency to reach out to customers using small-scale systems, trying to build out some install base with small-scale use cases, which are never put in production. You look like a more quiet company for that respect, focusing on developing the technology. Am I right or wrong in this interpretation? Is there a kind of explicit choice to be focused on the technology and not be distracted by what investors are asking to do, asking many companies to do, which is reaching out to customers, create a use base, and so on.

Chris Ballance: we're very strategic in this. We work with some real key customers, and we do this in a very selfish way. We want to learn how to build truly valuable systems. And for that, you need a few really key customers who know exactly what they're going to do, have very clear roadmaps generating value. know how to do that. And you get that from three good customers. You don't need 300 people doing this. And our big focus is on scaling out to build truly useful systems. And we see some really interesting use cases starting to turn on. If you just have a few hundred qubits with 10^{-4} errors, that's when some of the early chemistry applications, material science applications start to turn on. If you also want to really start properly learning about quantum error correction, that same size scale is where you can start not just dealing with one or two logical qubits, but dealing with many tens of logical qubits and properly learn about the intricacies of engineering these systems. So that's our big first, getting out to those products. And lots of the customers we're working on are really interested and well aligned in scaling out to that and deploying systems that rapidly get to that size scale.

Olivier Ezratty: so it means you believe that you can implement early chemistry solutions with a NISQ ploy, with variational algorithms like VQE, but thanks to the qubits you have, you're going to have maybe a larger circuit size. more than a couple of Ks of gates and stuff like that. So with or without the error mitigation, I presume.

Chris Ballance: that's exactly right. Importantly, with the very low errors we have, and also with things like some error mitigation, but also some symmetry protection by using a few tricks like mid-circuit measurement on different parts of the circuit. When you put all of these things together, you can start doing some really exciting stuff on these few hundred qubit systems. And importantly, at that point, you're well beyond the range where you can do any sort of modeling of these kind of things on a classical computer.

So you're really at the point where to start testing out and validating and using these applications, you really need proper quantum hardware. And importantly for us, all the problems we have to do to deploy these few hundred qubit systems look very similar to the problems we have to do to deploy the few hundred thousand qubit systems. So at that point, you're again doing more of the same to scale. The problems change, the flavors change, but a lot of the technology, logical learnings are very well aligned. And this is our focus. We don't want to get distracted by sidesteps on the route of very large-scale, both tolerant quantum computing.

Olivier Ezratty: a very practical question: how many are you in the company, right.

Chris Ballance: we're about 80 people now. We have our headquarters based in Oxford in the UK and then a moderate team in Colorado in the US and a small team in Switzerland.

Olivier Ezratty: what brought you in Colorado? I presume some trapped ion skills there. There's NIST and a lot of skills there. And one competitor, by the way.

Chris Ballance: there's two places in the world, two academic groups in the world that really worked out how to develop this electronic qubit control. One was based in Oxford that I led and one that was based in Colorado, led by David Wineland and David Allcock over there. And David Allcock now leads our US office based in Boulder.

Olivier Ezratty: and it's in Boulder. There are other teams in Europe and elsewhere, but the largest teams that you mentioned are the right ones indeed.

Chris Ballance: they've worked out how to solve some of the fundamental problems and really get those fidelities up to world-leading standards.

Olivier Ezratty: it was an exciting discussion. I'm excited to see that it may be possible to develop affordable systems. That's very important because I would like to avoid that all the systems cost billions of dollars because otherwise the market would be very small. I hope you're going to be an enabler of not a volume market like for PC's market, but at least a volume market for computing and scientific computing and chemistry is one of the key use cases that you mentioned for those companies. So thank you so much, Chris, for all these explanations.

Chris Ballance: absolutely. It's been an absolute pleasure. Bye-bye.

*PS: Fanny Bouton and I have been hosting the **Decode Quantum** podcast series since 2020. We do this pro-bono, without an economic model. This is not our main activity. Fanny Bouton and I are active in the ecosystem in several ways: she is the "quantum lead" at OVHcloud and cofounder of the France Quantum event, and I am an author, teacher (EPITA, CentraleSupélec, ENS Paris Saclay, etc.), trainer, independent researcher, technical expert with various organizations (Bpifrance, the ANR, the French Academy of Technologies, etc.) and also a cofounder of the Quantum Energy Initiative.*

Cet article a été publié le 15 avril 2025 et édité en PDF le 15 avril 2025.
(cc) Olivier Ezratty – "Opinions Libres" – <https://www.oezratty.net>