Topic 2d - The scale of cloud-based processing and storage

I'm at the UK National Center for Atmospheric Research, and I'm based at the University of Reading. And my particular interest is climate science. But I'm also interested in wider environmental science problems. And so for me, JASMIN enables us to do things with a scale of data we couldn't do before.

A key part of climate is doing projections for the future. A key part of trusting projections of the future is confronting our models with real-live data from space and from ground. And here at JASMIN, we have all of those things in one place. We have climate projections. We have data from space. We have radar data. We have people's observations from their backyards. And we can bring them all together into a melting pot. And that's unique worldwide.

JASMIN's a super-data computer. We really struggled to find a good name for it. It's a supercomputer. It's a data cluster. It's all about data.

We designed JASMIN to support the academic community in doing new world class science-- the academic community being really the environmental scientists in the UK, but also globally. And so we're looking to support people who have problems that are very small, but want to share their data-- and people who have got very large problems, who can't do it anywhere else. And we're also trying to help support industry, because we see this very much as pump priming-- a way to get from academic knowledge to industry.

So we're here at the JASMIN facility, which supports UK environmental science and earth observation. We're about to go in and look at the disk storage that we've purchased in the last year. And Jonathan's going to show us some of the new kit, so we can find out how we can use it in the future.

Sure.

OK.

So we have a mix of the very latest storage here. This is our object storage platform-- two racks, five petabytes. It's amazing to think that just these two racks contain more than the information that's at the far end of this aisle, which is JASMIN Phase 1, from 2012.

The rest of this aisle is all computer servers and storage. So these racks here are all from this year's purchase. We have about 10 or 12 racks-- about 45 petabytes of usable storage. And that's compared to about four or five down the far end.

So this is our original storage. Actually, we still bought some more of it today. We have the largest installation in the world of this kind of that particular storage. That's a parallel-storage platform composed of plates. Very easy to manage, which is why we originally bought it.

So here's an example of our computer. JASMIN is mostly storage. As I said, about 50 petabytes right now of usable storage, but only very small amounts of compute. So we're now at about 10,000 cores, which many years ago used to be a big thing, but today it's quite small. But that's why JASMIN is a super-data cluster.

Here, we have about 48 servers in a rack. The key thing is the networking out of the bottom. So those orange fibers down the bottom, each one of those is a 40 gigabit fiber. There are 24 of those coming out of the back of the rack. If you think, just from that rack alone, you're talking about 100,000 broadband connections-- just from this rack alone. So we can move data in and out of the rack at full rate, without ever getting a network problem.

So the network's a really special part of JASMIN. It's crucial to our success-- the ability to move lots of data from storage to compute. And it is the custom-designed network that makes it really special.

Yeah. So the net result of pushing all of that data through means our computers actually run really hot. This can get up to 40 degrees coming out of the back of the rack. So standing here is not a lot of fun.

Pretty hot here right now.

We now move on to JASMIN Phase 2 and 3. So everything you see in this aisle-- there are more than 10 racks on their side. And everything you see pretty much down the far end is all JASMIN. Again, in phases, so this is JASMIN Phase 2-- which was 2013-14. We have JASMIN Phase 3, and then what we call 3.5.

Different kinds of storage, so these racks here contain the block storage, different vendors again.

So the block storage is a key part of our support for the cloud. So we have our own private cloud in JASMIN. And this is the bit that makes the computers in the cloud work.

So some of the servers we have are different than the ones I just showed you here. These servers in here are designed to do very high memory. So each one of those core servers has two terabytes of memory in it. And that allows us to do very large problems in a single machine. The other computers are much smaller. And if you could do distributed memory, those would work really well.

So these are for the problems where we have big mathematical matrices that we have to manipulate, and we can't fit them on the smaller computers.

The four racks here are distributed storage. Those four racks are from about 2014-15. And those contain about four petabytes of data. That's the equivalent of the 12 racks you've seen over there. So now we've gone from 12 racks to four racks here. And if you remember the two racks I showed you right in the beginning, that was two racks with five petabytes.

So that's showing you the migration of the size of disk drives over that time period. Actually, a very short period of time from three terabyte drives all the way through to 12 terabyte drives today. So and increase of four in six years.

So in these racks, we have the very latest. These have the 12 terabyte drives. Again, they're distributed storage, but it's a different kind of storage. And one of the key things about JASMIN is, just like the tape, we have different kinds of storage to work with different workflows. So this is high-performance storage. This is super-high-performance storage, but very high performance in a smaller area. And then we have larger things, but slower, and different accesses for different people.

It turns out that when you have millions and millions of files, you need a different kind of storage for when you have lots of very large files. So these ones are for the very large files. And this is for the millions of very small files.

Yeah, that's right. So I'm only going to show you a very brief glimpse of the heart of the networking. So in these two racks here, we have the heart of the JASMIN 4 network. In order to stitch all of that infrastructure into all of the new things that you see here, we had to build a completely new data-center-level network. So instead of three terabytes a second, we're now sort of 24 terabytes a second, to get all of the things together.

It's a very unusual design. It's only the very big guys-- the Facebooks and the Googles-- who have networks of this kind. This is the top of the stack. The mocha fibers are 100 gigabit, instead of the orange ones we saw that were 40 before. And again, that's showing the step up in technologies in the last six years. So 100 gigabit is in just one fiber. Instead of the 12 fibers I showed you before, we can now do the same thing in one fiber, for all those hundreds of broadband connections.

So the key thing about this is, if you look at the number of 100-gigabit cables, you could almost imagine that the whole of the nation's broadband capacity. It probably isn't. But we're getting up there.

Yeah, that's right. In order to stitch those together, we have a huge network. Again, that's the top of the JASMIN 4. This is the very top of the tree. And so these 16 switches here, the reason that they're unoccupied is we're going to build on this as we go forward. And this includes links to other data centers across on site-- again, at multi-hundred gigabit.

And for that, we've had to provision something like 64 100-gigabit fibers between this data center and the one across the road. And we'll do that as we build out, just to handle the data rate.

So the storage that we've got right now, the combination of the new storage here and the old stuff, will bring us in at about 500 gigabytes a second. Which doesn't sound very much like a particularly big number, but when you think that would place us probably in about number three in the world for I/O performance, it's pretty impressive. Even the world's number one is only four times bigger than that.

So if you imagine that most of this whole infrastructure here is $24 million over six years-- something like that. And each part of that has to be procured through the team here. Part of that is myself. But, obviously, we have some other people in the organization working on it as well. So it is a big problem.

But that's actually driven what we're seeing, with the different kinds of storage now. We used to just have one kind of storage, which would be good enough for everybody-- but clearly, it wasn't ideal. Now, we've got multiple kinds of storage.

But the reason this network is key here is because this is the thing that's allowed us to stitch those disparate bits of infrastructure together in a way that allows us to kind of plug and play. So we can take this part of infrastructure. We can bring it with the other part. There are no boundaries between any parts of those infrastructure. And that's really almost unique.

Yeah. This is the equivalent in hardware terms of what's happened in the software industry in the last decade, where they talk about agile software development. We have an agile hardware development. And Jonathan and Steve, every year, come up with new ways of putting things together. They add new functionality and new performance, so we can get up there and do things we couldn't do before.

Yeah. I only have a team of two people working with me for all of us. This is currently 70 racks of kit. We clearly didn't install all of this. But we have to maintain it. We have to manage all the connections. If you think if just the number of network connections in here, and the number of individual IP addresses-- so we're probably about 10,000 IP addresses now. If you think, your broadband connection at home only has one of those.

So we have 10,000 of those endpoints. And the number of network connections-- I don't know, we have a massive database of these things. It's, again, in the tens of thousands of individual cables. And if you look, each cable has a label on it, which tells you where the endpoints are. And each one of those endpoints is in the database, so we know where it is, what's gone wrong. We have monitoring for all of this stuff.

So we're here with our tape library. A lot of people think that tape is an old technology. It is indeed an old technology, but it's got more future than solid disks.

So in here, we've got some robots that are moving the tapes in response to user requests. So we have up to 10,000 tapes in this. And there's another one here with another 10,000 in it.

And this is how we used to do things. It used to be users would request some data. And they would get some data, and the robot would bring them a little bit of data. And they'd get on and look at their little bit of data.

Now, we can do things with massive amounts of disks, but the data still has to start from the tape library. And when we can't fit it on the disk anymore, it has to go back to the tape library.

And if we look forward five or 10 years, tape will be incredibly important, yet again. Because the future for the way that they're actually building these things is much better for tape than it is for disk. And you can see a little robot's just moved there, and it's grabbing a tape. And it'll be loading some data and moving it to disk for some real users.