Microsoft Research Faculty Summit
Tony Hey welcome
Redmond, Washington
July 12, 2010

TOM MCMAIL: And without further ado, I would like to introduce our Corporate Vice President, and the head of External Research, Mr. Tony Hey. (Applause.)

TONY HEY: Thanks very much. It's great to be here, and great to see all of you here in Seattle. So, I'm here to welcome you to Seattle, and it's great to see so many of you here, and so many new faces in the room.

We hope we've got an interesting program, so let me carry on and tell you a little bit about it. So, the goal of the summit is to network with colleagues and friends, and to understand the problems that are facing not only computer science, but also how computer science can help solve some of the problems of the world. So, we've got nearly 100 presenters, we've got 27 countries, 300 or so people who have not regularly been to the Faculty Summit, which I think is an impressive number, and I would be very interested to hear your reactions after the event, and from 170-odd institutions and agencies. So, it's a really diverse group, and I hope that you find it a useful exercise in networking and building relationships.

The theme for the conference is embracing complexity, which sums up the goal of computer science, which is handling complexity by abstraction, and hierarchical design, and so on. And we've picked out some of the major themes. There's a whole session on natural user interfaces; software engineering, of course; the future of the Web where Web is going; semantic Web; Web 3.0; Web 4.0; and then there's the cloud, which is actually going to change all sorts of things in industry, in the way we do things in research, and also for consumers.

And we're also seeing, and you'll find outside some copies of a book called, *The Fourth Paradigm*, which is a collection of essays about how the data explosion in most of the sciences is going to affect us all, and how I think techniques for computer science will be need to help scientists manage the challenge of this data to find the signals, and to understand the events that are happening.

So, addressing global challenges, the graphic is about the Ocean Observatory Initiative. There's one of them off Seattle in Puget Sound, putting a fiber optic cable on the seabed floor, which will change oceanography from being data poor, you just have a ship across the surface getting occasional bits of data to have data flooding in 24/7, 365 days a year. So, large amounts of challenges, and I hope will have a stimulating discussion. And we'll see exactly how we can engage and collaborate in these areas.

So, my organization is called internally External Research, but really it's about collaborating with universities doing collaborative research with universities, and we have broad geographical outreach, so we have visitors here from Asia, from India, from Europe, and Latin America, and each of these countries has an outreach program with universities, and of course they're all

slightly different in each of the countries, and we view this as very helpful, and we try and get commonality where we can, and diversity where it's appropriate.

We've organized our programs. Obviously we have a core computer science activity, and that's really a major focus for the future of our industry. We believe that it's important that computer science departments continue to thrive, continue to produce great students, and that the students are interested and excited to do computer science. And part of the way to do that is actually by showing how computer science can help solve some of the problems of the world. And earth energy, environment is one of the themes, and you saw the opening video about one of the projects we have in Latin America with the sensor network in the rainforest. Health and well-being is another major strand of our activity where we're applying to the wealth of genomics data, for example, machine-learning techniques that are going to be essential to make understanding of that, and to understand complex diseases like HIV-AIDS.

Lastly, in addition to this data revolution, and the ongoing revolution in computer science, we are in the midst of a publishing revolution when you can make free digital copies for nothing, and when you can actually publish it on the Web yourself for nothing, that really changes the game as newspapers and everything are finding. And so, what does that mean for scholarly publishing, and how can we actually link the data to the publications, and what are the implications of that, and how do we store data so we can access it in 20 years time, for example. So, education and scholarly communication are an important theme, and you'll find that's one of the themes at the conference, too.

And the rest of my team, we try and work with the university community to try and build tools, and software services which actually help them do their research. And so, the goal is to work with researchers to try and product tools, and at the moment we're doing them in open source so we can extend them, you can extend them, you can port them to other platforms. So, that's what we're trying to do. These are our major disciplines, and this is where we focus our efforts.

So, what I would like to do is take an example of how computer science, and the technologies from computer science, can actually affect some of the other disciplines. So, what we're announcing today is the Terapixel Project, which is available today on WorldWide Telescope. And, I hope you can see here, this is the original image WorldWide Telescope uses of the sky taken from digitized photographs of the sky which produces overlapping regions, they're taking in slightly different exposures, and so on, and the edges are slightly untidy. And it's around three million files, and about half a petabyte of data. And so what we wanted to do was bring them all together and actually the 15-year set of plates could remake them into a seamless image that actually looks like the actual sky that you see.

So, in the moment, state of the art is this individual tiles and seams, and what the challenge was, can we manage and manipulate the vast amount of data, first of all to take from the raw data to create the color image, then you have to stitch and smooth them, and then you create the sky image pyramid that you can zoom in and zoom out for WorldWide Telescope. So, that was the challenge.

And to do that, what we needed to do was put together a bunch of computer science technologies. So, on the left you see creating the color plates, and the various tasks you have to do there. Then stitch and smooth the images using distributed gradient domain processing to smooth them, and then finally to create this pyramid. And the technologies that we use to do this—it was sort of a six-month challenge. And it was, I would say, a grand challenge to produce an image this size, which is large numbers of HD TVs with hundreds of thousands of them would have to accommodate this image.

We used Live Link, which is a type of map reduce, a slightly more general version of map reduce. We used the .NET parallel extensions, because we needed to do parallel processing to decompress the data, and then we used high performance computing on a Windows cluster, with Drive Link. And the whole thing was orchestrated by a workflow engine called Trident that we produced. And putting all these different technologies together they were able to process these different technologies together they were able to process these 3 million files and produce a really impressive image at the end of the day.

So, I hope you can see a difference between these two images. It's not that we failed. But, that was what it was like, and the final image is really very impressive and 500,000 high-definition television sets it tells me you need to view this. So, it's a large amount of data, and you can access it from your desktop and zoom in and zoom out. So, I do encourage you go to and try WorldWide telescope with this new data and get lost in space. So, it's really good.

So, that's Terapixel. What I'd like to do now is announce the official release, the Apogee release, of WorldWide telescope. It's a new release. It has all sorts of cool things, exciting things like the asteroid belt, and you can do amazing things with it. And in just a second I'll introduce NASA's Chief Technology Officer for IT, Chris Kemp, with whom we collaborate with NASA on taking their Mars data, but first of all I'd like to show you a brief video. So, could we roll the video please?

(Video segment and applause.)

Thanks, and now Chris will tell you really about it.

Chris.

CHRIS KEMP: Hi, thanks today. We're really excited about the release of WorldWide Telescope Mars. This was a really exciting project. We started working with Microsoft research about a year and a half ago. And teams from NASA that were responsible for doing all of our data processing for our lunar data sets were looking at ways to make that data more accessible on platforms that had more reach to the public. So, one of the things that we wanted to do is ensure we were able to connect with classrooms, universities, on a platform that was extensible and accepted by millions of people worldwide. So, what you're able to see today in WorldWide Telescope when you zoom into Mars is the largest digital image mosaic ever created.

The technical challenges for this project were immense. We have almost 15,000 terapixel images, over a billion pixels per image, each of these images is, of course, too large to download.

So, we worked with the Microsoft team to develop a high-resolution projection format, unlike those in other projects like NASA's own World Wind project, and other commercially available geospatial browsing platforms. The Microsoft WorldWide Telescope platform does not break down at the poles, and the poles are some of the most interesting regions on Mars. So, we saw this as an exciting opportunity to create a platform that was much more scientifically accurate and useful as we look at some of the more interesting regions on the planet.

We developed our image-processing pipeline to run in the cloud. So, we were able to take these 15,000 terapixel images and turn them into almost half-a-billion smaller PNG images that are progressively downloaded when you connect to the product, and you go to an area. So, as you're zooming in you're actually only downloading the image content that you need to see whatever particular region of the planet is on your screen. The other really interesting technical challenge was applying extremely high-resolution terrain maps to the surface. No one has ever really zoomed in as far as we were allowing folks to zoom in. So, as we zoomed in, we actually had to work really closely with the Microsoft team to optimize some of the 3D algorithms in Silverlight to render the ultra high-resolution terrain maps without any visual image artifacts.

So, this was a great example of a public-private partnership that resulted in a massive data set that was always available on the Web if you wanted to download terapixel scale images from a JPL website. But, by making all this content accessible on WorldWide Telescope, we're making it accessible to classrooms, to labs, to scientists that want to build on top of that existing platform and I really see this as the future of NASA's data architecture. Whereas, instead of focusing on building end user applications the government is a platform and we're working with organizations like Microsoft Research to develop standards to deliver the data so that it can be consumed by the scientific community.

So, it's been a fantastic collaboration. I appreciate the opportunity to work with Tony and his team, and there's more to come. We have lots of data.

TONY HEY: Thanks very much, Chris. (Applause.)

It's a very exciting collaboration for us. We have a NASA Space Act agreement. And we hope there's lots more good things to come. So, it's a great collaboration.

So, that's applying computer science to solve some of the problems of, in this case, displaying astronomy data. What about computer science for computer science sake? Well, what we've also got is a service called Web N-Gram, it's now in public beta, and its intent is to advance research in search and what goes beyond the search, and beyond search, and you'll hear about that in one of the parallel sessions today.

To do that researchers need access to real world, large-scale data and to be able to do data and compute on this data. So, what we're trying to do, we've been struggling with the problem of the privacy and anonymity of data that are critical, and therefore what we're trying to do is make this available to the research community, access to Web scale data via data services.

So, this is the Web N-gram service, which is to help advance research and experimentation in all sorts of areas, such as speech, learning, machine translation, as well as search. And so what's it about? Well, it's about this: users usually have a query on a single word, and the last—the blue lines there show single word diagram, trigram, and fourgram distribution. But, search engines typically only search on unigrams and so the distribution you get from the search engine is not the ideal distribution that you would normally get. And that's partly because there's more information than just the body of the text, or the webpage, there's the title and the anchor, and all this context metadata is typically ignored. So, this will be the subject of a workshop, as you see there in Geneva, later this month, and those of you who can have the opportunity to go to Geneva, I like the town very much. I used to live there. So, I do recommend that you go there. And also the conference will be interesting.

So, let me just give you an example. What was really good, when we announced is at the last WorldWide Web conference in Raleigh, and within eight hours Dr. Ding at RPI had produced this multiword application.

So, you can see on the left a single tagged cloud and you see the usual from a body of government data he was using and you see it picks out a number of tags. But, you get much more information if you use the engrams. So, you see critical and habitat, but you see that it comes together in critical habitat, and you see also things like toxic release inventory data is a fourgram, and that's extra information that can actually help make clear the user intent much more than just searching on single words.

So, large amounts of things to be done, this is free for non-commercial research. We have a petabyte of data there in the cloud, and it's assisted by our ISRC team, the Internet Services Research Center. And they've been very helpful and you can collaborate with them. And it's available on Azure for people who are participating in the NSF program for computing in the cloud. It's a collaboration with Bing and Harry Shum will be giving a talk later today about the collaboration with Microsoft Research. So, that's an example of actually doing something for the computer science community. So, we don't just do things for the environmental science and health communities, but computer science is an important area for us.

So, where are we? Well, I'm standing in the way between you and an interesting program. So, we are going to announce tomorrow the new faculty fellows for this year, and we have a little presentation with Rick Rashid and I invite you all to that. Today we have a number of exciting plenaries and themes, sessions. So, we have themes in Natural User Interface, Data-Driven Software Engineering, Challenge of Large Data, Future Web, and there's Design Expo. And tomorrow we have Operating in the Cloud, and Demofest, as well as some special topics, and the special topics are also extremely interesting, in my view. I have difficulty deciding which ones I can dip into. But the plenaries I hope you'll all enjoy. So, the first plenary we're coming up to in a moment is the innovation of what used to be called Natal. It's a great collaboration between the Xbox team, and Microsoft Research. Several of our labs participated in that, and you'll hear about that in the next talk. Then we have this year's Turing Award Winner Chuck Thacker, who essentially was one of the pioneers at Xerox PARC for the infrastructure you see around you, the PC with the mouse and the keyboard, and the graphic screen, and so on. That was all developed

by the Alto team, of which Chuck was one of the major leaders, along with people like Butler Lampson.

So, Chuck will be giving talk, Rethinking Architectural Research and Education. And then tomorrow morning we start with a panel chaired by Rich DeMillo, with people like our very own Ed Lazowska, who is sitting in the front row, who will be there tomorrow morning along with others from around the world.

And, finally, we finish with what I hope will be an exciting talk about the making of Avatar, and I assume most of you have now seen Avatar in 3D. The interesting thing is whether, Justin Rattner gave a talk at Supercomputing last year, and it was about the 3D Internet, and he clearly sees that the 3D is a major part of our future, and that's what Hollywood thinks, so it will be interesting to see exactly what this audience thinks of that.

If you want to tweet, that's the symbol you'll use. There's a dinner cruise tonight, and I hope that you'll enjoy the networks, and be able to leverage the space. We've tried to make it a little less structured than last time, so you have an extra amount of time which is not taken up and organized for you, you can use for whatever purpose you like. So, I hope you have a great time.

And what I would like to do is finish here, and introduce our next speaker, who is this year we're focusing in our computer science theme on software engineering, and programming languages, and Judith Bishop is leading that activity, but we have a new activity on natural user interfaces which is being led by Chris Tull. So, thank you very much for listening, and I'll introduce Chris Tull in just a moment. Thank you. (Applause.)

END