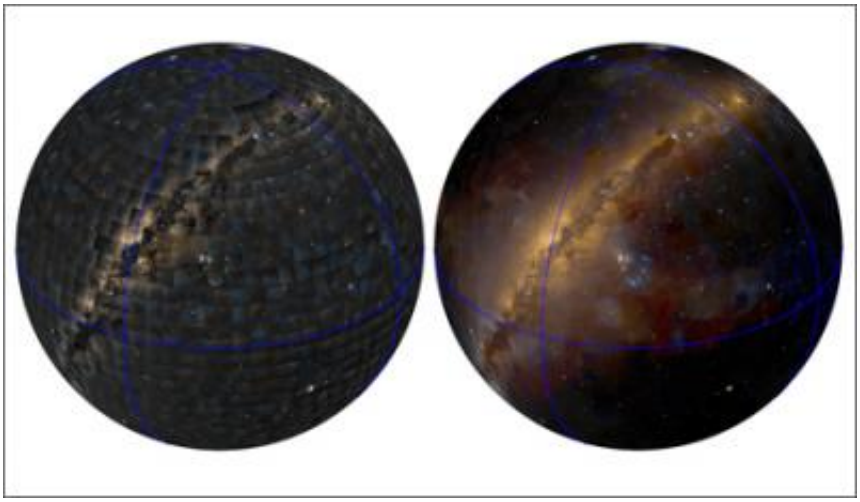
Terapixel Project: Lots of Data, Expertise

July 19 2010, By Rob Knies



A view of the Milky Way (left) before Terapixel processing displays tiled artifacts, but the post-processing image is virtually seamless.

How can you achieve the impossible? Easy -- as long as you have the right people and the right tools. The Terapixel project from Microsoft Research Redmond is proof positive.

The effort—to create the largest, seamless spherical image ever made of the night sky—has tantalized astronomers for decades, but the sheer volume of data and the challenges in data manipulation have proved frustrating.

Until now, that is. A small but energetic team from the External Research division of Microsoft Research has found a way to use a

collection of Microsoft technologies to produce the largest, clearest sky image ever assembled. The project was unveiled July 12, the opening day of Microsoft Research's 11th annual Faculty Summit.

The image, available on both Microsoft Research's WorldWide Telescope and on Bing Maps, surpasses the gargantuan size of 1,000,0002 pixels—one terapixel. Dan Fay, director of Earth, Energy, and Environment for External Research, says that to view every pixel of the image, you'd need a half-million high-definition televisions. Alternatively, it you were to attempt the print the image, the document would extend the length of a football field.

And it's not just big. It's effective, too. Astronomers who have received an early look at the image have been astounded.

"It is absolutely gorgeous," says Brian McLean, an astronomer at the Space Telescope Science Institute who supplied the original images for the project. "It is now truly a seamless all-sky image. As someone who has worked on the creation and processing of the [Digitized Sky Survey] for the last 25 years—using it for both science and telescope operations—I can appreciate how the application of the new workflow and high-performance-computing technology has made this possible."

Roy Williams of the California Institute of Technology also has worked with the researchers involved in the Terapixel project, and he, too, is impressed with the results.

"You've done a fabulous job making the new all-sky mosaic. … For the first time the full glory of the original data is revealed. The new image layer is a real improvement in clarity and beauty, done with smart algorithms and a lot of computing!"

In May 2008, when the WorldWide Telescope launched, bringing an

interactive visualization of the heavens to anybody with a computer, the Digitized Sky Survey images were used. Like many panoramas, however, the stitched collection included artifacts from varying exposure levels across the individual parts of the panorama. There were differences in brightness levels and color saturation, as well, and while you could pan around and zoom into mind-boggling galactic imagery, the panorama was not seamless. The varying exposure levels led to edges that didn't match up, and the resultant stitchings sometimes looked like a checkerboard.

That didn't please Jonathan Fay, a principal software architect who works with his namesake, Dan. (The two are unrelated.)

"The original vision of creating this seamless stitch was something I wanted to do myself for a very long time," Jonathan Fay says. "Eventually, it had to be put on the back burner, because it was so much work and we didn't have the infrastructure to do it. This is something I was really passionate about, but I knew getting the resources to do this would be challenging."

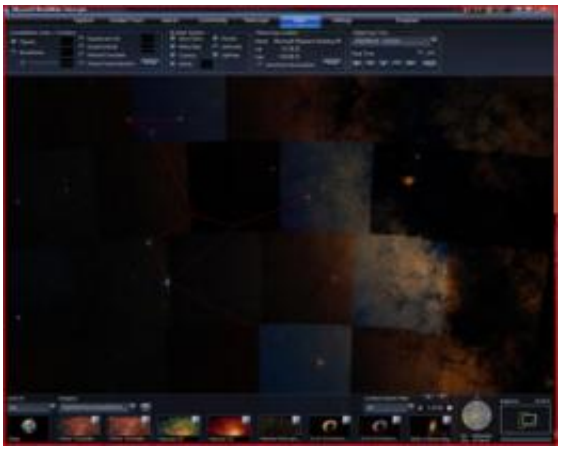
In December 2009, however, the two Fays put their heads together.

"We were trying to think of things that the ARTS [Advanced Research Tools and Services] team could do, work that is helpful to WorldWide Telescope but somewhat relatively independent," Jonathan Fay recalls. "I shared with them this passion to finish this and do all our sky panoramas as seamless stitches. It sounded like the Trident workflow was something that could be useful for it. That kind of resounded."

It certainly did with Dean Guo, senior program manager for the ARTS team. He, along with lead developer Christophe Poulain, swung into action.

“Christophe and I were interested in large-data-set computation and processing,” Guo says. “We were looking for a showcase study. This one came along, and we decided, ‘OK, this looks very interesting,’ so we signed up for that. But we didn’t know what we had signed up for.”

In the end, though, they got the study they were seeking. “There is massive computation involved,” Guo says.



A view of the constellation Sagittarius, before Terapixel processing ...

... and after the Terapixel smoothing.

“It’s a huge endeavor,” Dan Fay stipulates. “The challenge would be how much data and how much space we would have and how to move the data around.”

The Digitized Sky Survey produces photographic plates of overlapping regions of the sky, using images collected from a pair of telescopes, the Palomar Observatory in San Diego County, Calif., and the Anglo-Australian Telescope in the Australian state of New South Wales. Over a period of 50 years, the two telescopes combined had captured 1,791 pairs of red-light and blue-light images that covered Earth’s entire night sky. Those images were scanned over a 15-year period into a series of plates, compressed and stored as tiles. The result was a collection of 3,120,100 files that, even when 10-fold compression was applied, took up 417 gigabytes of storage.

How could a small team cope with such voluminous data? Not easily, and probably not at all a few years ago. But that was before Project Trident, a scientific-workflow workbench that debuted in the summer of 2009. Trident makes complex data visually manageable, increasing the scale at which scientific exploration can be conducted.

“Part of the challenge,” Dan Fay says, “is just the sheer magnitude of the data. It had to be distributed across machines to process, to create an image pipeline. That’s where the benefit of the Trident workflow came in. It enabled Dean and Christophe to rerun the data multiple times so we could improve the quality of the image and the smoothness of it.”

Guo and Poulain used Trident workflows and DryadLINQ to manage code running in parallel across a Windows High Performance (HPC) cluster. All of this was enabled by the researchers’ access to the Microsoft Research Shared Computing Infrastructure, and between

faster program execution delivered by the .NET parallel extensions on multicore machines and processing the data on a 64-node cluster, the time it took to run the job shrunk from literally weeks to a few hours.

Other issues appeared, as well, such as vignetting, which is a darkening of the edges and the corners of each plate. Each plate required correction to contribute to a clear, seamless image, and Dinoj Surendran, data curator for the project, proved invaluable in this effort.

The stitching and smoothing process also presented challenges, particularly in adapting the telescope data to a spherical model to avoid the distortions near the poles common to two-dimensional maps. To address these concerns, the project team worked with Hugues Hoppe and Michael “Misha” Kazhdan.

For a few years, Hoppe, a principal researcher and manager of the Computer Graphics Group at Microsoft Research, had worked on various projects with Kazhdan, a professor in the Department of Computer Science at Johns Hopkins University. In 2008, they wrote the paper Streaming Multigrid for Gradient-Domain Processing on Large Images, which showed how to assemble photographs efficiently to form seamless gigapixel panoramas. Then in 2010, along with Surendran, they published the paper Distributed Gradient-Domain Processing of Planar and Spherical Images. Their contribution was to generalize seamless panoramic stitching to the case of spherical images and to make it scalable on even larger data.

“We had a big optimization to do,” Hoppe says, “with lots of unknowns. The unknowns are the pixel colors. We’re trying to solve for the pixel colors of the stitched montage, such that all the seams will disappear. The constraints are that the neighboring pixels should relate very much like in the original images, and at the seams, where you have different images merging together, the colors should be almost identical so you

don't perceive any discontinuity.

"It's a giant optimization problem, and our approach is about making that efficient. There are many heuristic techniques that people have used before. That helps attenuate the seams, but it doesn't fix them correctly."

Hoppe and Kazhdan turned to Poisson image editing, the result of work performed at Microsoft Research Cambridge in 2003.

"Our work is similar to that," Hoppe says, "but it's about making it efficient in very large domains. Our initial work in 2008 demonstrated results on gigapixel panoramas, and we were impressed that we were getting results there. Later that year, we went up to the terapixel, a thousand times larger."

The mapping of the sphere was handled in part by TOAST, for Tessellated Octahedral Adaptive Subdivision Transform, the spherical projection system used in the WorldWide Telescope.

"This particular parameterization, it's all very nice and continuous, except you have to deal with the boundary conditions," Hoppe says. "What's tricky is that one edge of the planar domain has to be a mirror image of another, and that's true across all of the boundaries. If you try to solve for a seamless image not respecting these constraints, you would see these discontinuities across four of the edges."

Plate data is transformed into a grid of tiles. The grid is divided into columns, called strips, and each is sent to one HPC node for processing in parallel. The data is distributed over a cluster, necessary because of the massive amounts of data involved.

"These difficult boundary conditions," Hoppe explains, "are handled by the fact that they're local in the memory."

As the strips are processed, the nodes talk to each other.

“If we were just to do the optimization separately,” Hoppe says, “it would create seams, because the optimization results wouldn’t be interacting. We have communication that’s happening between these nodes over the network of the computer cluster.”

Guo, Poulain, and their development agency, Aditi, also found themselves reliant on Jonathan Fay, chief architect for the WorldWide Telescope project and an avid astronomer.

“When we started the project, we knew nothing about astronomy,” Guo smiles. “Our domain knowledge was really limited. But during the project, we learned quite a bit.”

Fay, of course, was happy to help.

“They really dug into this,” he says. “The more they dug into it, the more they wanted to know. We realized they were going to have to be bootstrapped on astronomy concepts, coordinate spaces, the Digital Sky Survey—all these astronomy terms that I took for granted. I would sketch out the architecture on the whiteboard, and they would work on it and ask me refining questions.

“They put the sweat equity in to make this happen. I just used the astronomical knowledge in my head and my image-processing knowledge to give them some course correction and guidance.”

That effort includes all the work to provide the ability to zoom into and out of the galaxies the image data represents. The sky-image pyramid includes more than 16 million tiles, each one 256 pixels square. Such scale—and the scientific opportunities it offers—is the key concept behind the recent book The Fourth Paradigm: Data-Intensive Scientific

Discovery.

In fact, several individuals associated with the Terapixel project suggest that the effort could serve as a model for other work involving big data for scientific exploration.

“The parallel extensions, how you use the workflow to manage the clusters … that’s not just limited to the Terapixel project,” Guo says. “It could be any large, data-intensive, and computationally intensive project.”

Jonathan Fay agrees wholeheartedly.

“This is stunning evidence of what can be done with Trident and clustered, high-performance computing, and Dryad,” he says. ”There are probably a lot of stories of people who have said, ‘I’ve got a bunch of data that needs a bunch of processing with a sophisticated pipeline.’ There are a lot of those stories out there, and they’re not just astronomy-image-processing issues.”

Thus, the Terapixel project remains a cutting-edge research effort.

“I’ve shown this to a few other astronomers,” Fay says, “and they all think this is absolutely beautiful.

“I’m really indebted to Dean and Christophe and the whole crew,” he adds. “I know what a phenomenally difficult project this, and while Trident made it a whole easier, it didn’t make it free. They still had to do a lot of work themselves. I really appreciate what they’ve put into it.”

So, apparently, do others, including Tony Hey, corporate vice president of External Research.

“This was a tour de force,” Hey says of the Terapixel effort. “Can we make the tour de force into the routine? That’s the challenge now for my team.”

Something suggests that such a challenge might be met successfully.

“In a way, it’s why you come to work at Microsoft,” Poulain says. “You get to work on very interesting problems using state-of-the-art solutions, and you create a solution that millions of people potentially will be able to look at. That’s pretty awesome.”

Guo concurs.

“I can’t imagine we could have solved this problem two years ago,” he says. “With the new technologies available, a terapixel is no longer a big number, so we are able to solve this problem today. It’s very exciting.

“Who knows? Maybe five years from now, this project will look small. At Microsoft Research, we are able to do things like this to really showcase what we’re capable of.”

More information: research.microsoft.com/en-us/p … rapixel/default.aspx

Source: Microsoft

Citation: Terapixel Project: Lots of Data, Expertise (2010, July 19) retrieved 24 April 2024 from https://phys.org/news/2010-07-terapixel-lots-expertise.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is

provided for information purposes only.