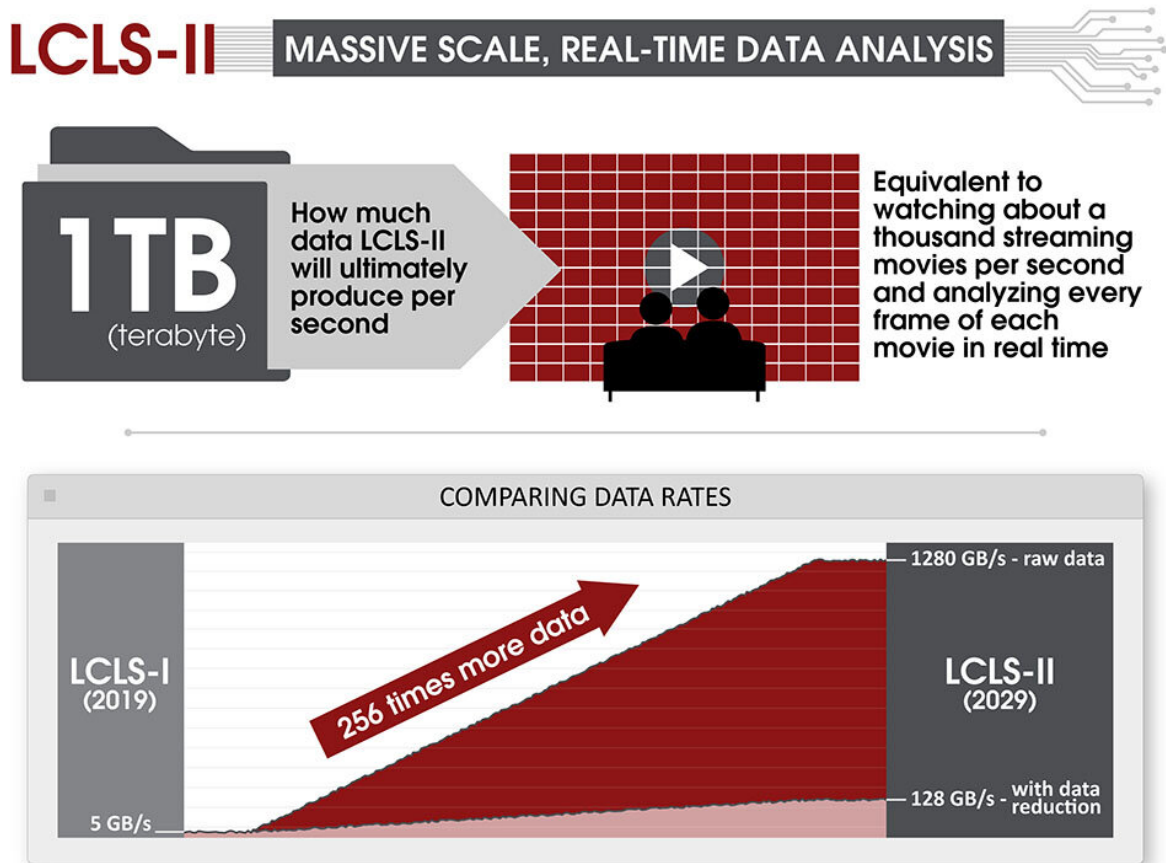# SLAC's new X-ray laser data system will process a million images a second

February 18 2021, by Niba Audrey Nirmal

Credit: Greg Stewart/SLAC National Accelerator Laboratory

When upgrades to the X-ray laser at the Department of Energy's SLAC

National Accelerator Laboratory are complete, the powerful new machine will capture up to 1 terabyte of data per second; that's a data rate equivalent to streaming about one thousand full-length movies in just a single second, and analyzing every frame of each movie as they zoom past in this super-fast-forward mode.

Data experts at the lab are finding ways to handle this massive amount of information as the Linac Coherent Light Source (LCLS) upgrades come on line over the next several years.

LCLS accelerates electrons to nearly the speed of light to generate extremely bright beams of X-rays. Those X-rays probe a sample such as a protein or a quantum material, and a detector captures a series of images that reveal the atomic motion of the sample in real time. By stringing together these images, chemists, biologists, and [materials scientists](link) can create molecular movies of events like how plants absorb sunlight, or how our drugs help fight disease.

As LCLS gets upgraded, scientists are moving from 120 pulses per second to up to 1 million pulses per second. That will create a 10,000 times brighter X-ray beam that will enable novel studies of systems that could not be studied before. But it will also come with an enormous data challenge: The X-ray laser will produce hundreds to thousands times more data per given time period than before.

To handle this data, a group of scientists led by LCLS Data Systems Division Director Jana Thayer is developing new computational tools, including computer algorithms and ways of connecting to supercomputers. Thayer's group uses a combination of computing, [data analysis](link) and machine learning to determine the patterns in X-ray images and then string together a molecular movie.

Going with the flow

At LCLS, the data is continuously flowing. "When scientists get access to run an experiment, it's either a 12-hour day or a 12-hour night, and limited to just a few shifts before the next team arrives," says Ryan Coffee, SLAC senior staff scientist. To make efficient use of valuable experimental time, bottlenecks must be completely avoided to preserve the flow of data and their analysis.

Streaming and storing the data presents a significant challenge for network and computing resources, and to be able to monitor data quality in near-real time means that the data need to be processed immediately. A vital step toward making this possible is to reduce the amount of data as much as possible before storing it for further analysis.

To enable this, Thayer's team has implemented on-the-fly data reduction using several kinds of compression to reduce the size of data recorded without affecting the quality of the science result. One form of compression, called veto, throws out unwanted data, such as pictures where the X-rays missed their target. Another, called feature extraction, saves only the information that is important scientifically, such as the location and brightness of a spot in an X-ray image.

"If we saved all the raw data, like we've been doing up till now, it would cost us a quarter of a billion dollars per year," Thayer says. "Our mission is to figure out how to reduce the data before we write it. One of the really neat, innovative parts of the new data system we developed is the data reduction pipeline, which removes irrelevant information and reduces the data that needs to be transferred and stored."

Coffee says, "Then you save a lot on power, but more importantly, you save on throughput. If you have to send the raw data through the network, you're going to completely overwhelm it trying to send out images every single microsecond."

The group also created an intermediary place to put the data before it goes to storage. Thayer explains, "We can't write to the storage directly, because if there is a glitch in the system, it has to pause and wait. Or if there's a network hiccup, then you can lose data altogether. So, we have a small but reliable buffer that we can write to; then we can move data onto permanent storage."

Driving innovation

Thayer emphasizes that the data system is built to provide researchers with the results of their work as promptly as the current system, so they get real-time information. It's also built to accommodate the expansion in LCLS science for the next 10 years. The big challenge is to keep up with the enormous jump in the data rate.

"If you imagine going from analyzing 120 pictures per second to 1 million per second, it requires a lot more scrolling," she says. "Computing is not magic—it still works the same way—we just increase the number of brains working on each of the pictures."

Supported by a recent award from the DOE, and working with colleagues from across the DOE national laboratory complex, the team is also looking to incorporate artificial intelligence and machine learning techniques to further reduce the amount of data to be processed, and to flag interesting features in the data as they arise.

To understand the LCLS data challenge, Coffee draws an analogy to self-driving cars: "They must compute in real time: they can't analyze a batch of images just recorded and then say "We predict you should have turned left on image number 10." SLAC's data rate is much higher than any of these cars will experience, but the problem is the same—researchers need to steer their experiment to find the most exciting destinations!"

The upgrades driving this massive leap in data rate and performance will come in two phases over the coming years, including LCLS-II and a high energy upgrade that follows. The work of the data experts will ensure that scientists can take full advantage of both. "Ultimately it will have a dramatic effect on the type of science we can do, opening up opportunities that are not possible today," Coffee says.

Provided by SLAC National Accelerator Laboratory