Preparing for tomorrow's big data

SPOTLIGHT | OCTOBER 2, 2013

Last week, the inaugural ISC Big Data conference was held in Heidelberg, Germany. The event was chaired by Sverre Jarp, chief technology officer of CERN openlab, and CERN was the focus of two case studies presented during the two-day conference. Frank Würthwein, from the University of California at San Diego, US, discussed how CERN handles big data today and looked forward to how the organization will have to adapt these processes to cope with increased peak data rates from the experiments on the Large Hadron Collider (LHC) after upgrade works are completed as part of the first long shutdown (LS1).

ISC Big Data '13 was held alongside ISC Cloud'13, which you can read about in this week's article 'Flexibility — for HPC, clouds, and the workforce'. Image courtesy ISC events.

Until recently, the large CERN experiments, ATLAS and CMS, owned and controlled the computing infrastructure they operated on in the US, and accessed data only when it was locally available on the hardware they operated. However, Würthwein explains, with data-taking rates set to increase dramatically by the end of LS1 in 2015, the current operational model is no longer viable to satisfy peak processing needs. Instead, he argues, large-scale processing centers need to be created dynamically to cope with spikes in demand. To this end, Würthwein and colleagues carried out a successful proof-of-concept study, in which the Gordon Supercomputer at the San Diego Supercomputer Center was dynamically and seamlessly integrated into the CMS production system to process a 125-terabyte data set.

CERN's Pierre Vande Vyvre also gave a presentation at the event in which he discussed the role of fast custom electronic devices in filtering out much of the data produced by scientific experiments such as those at CERN, so as to make the data more manageable. Currently, just 1% of data from collision events in the LHC is selected for analysis. "The big science workflows are mainly data reduction," says Vyvre. "The archiving of raw data is not the *de facto* standard anymore." He predicts that next-generation science experiments will reduce more and more the role of these custom devices and will instead entrust the processing of the complete data sets to standard computing algorithms. "One of the main problems faced by scientific experiments today is that lots of legacy software is in use that hasn't been designed for the big data paradigm," says Vyvre. "Tighter integration between data movement, scheduling and services is needed."

During his presentation, Vyvre also took as an example the Square Kilometre Array (SKA) telescope, which aims to have a total data throughput from its detectors of 10-500 terabytes per second by 2020. Yet, the conference didn't just focus on CERN, the SKA, or even scientific research in general. Big data use cases from a wide variety of industrial and business applications were also discussed, from modeling cars to improving postal services, and from optimizing marketing strategies to detecting online fraud. A prime example of work bringing together researchers from both science and industry is the the Helix Nebula initiative, which was presented at the event by Rupert Lück. The Third Helix Nebula General Assembly was also held in Heidelberg last week and Helix Nebula is now working towards production in 2014 (read more about the initiative in our recent article, '*View from above: a planet brimming with data*'). "What made this conference rather unique was the blending of speakers from academia and science with speakers from large enterprises, such as Google, Paypal, British Telecom, Virgin Insights, and LexisNexis," says Jarp. "All speakers described how big data can be used to create value from today's tsunamis of big data."

Another highlight of the conference was a keynote speech by Felix Wortmann, of the University of St. Gallen and the Bosch Internet of Things and Services Lab. He discussed the power of big data to drive disruptive innovations and, thus, have a lasting and wide-ranging impact. "Three years ago, big data was a topic for expert forums," says Wortmann. "But the notion of big data has now significantly changed." Jarp agrees: "Big data is serious: both software and hardware vendors are taking it seriously and are adapting their solutions accordingly. At CERN, we're gearing up towards being able to capture terabytes of data per second, but when it comes to 'the internet of things', this sort of rate could perhaps become commonplace."

- Andrew Purcell

Average:

Your rating: None Average: 4 (5 votes)

RELATED TERMS: analytics big data CERN Data management Europe grid computing ISC ISC Big Data biology cloud computing data management systems high-performance computing security and authentication standards information services workflow management systems portals, science gateways, and hubs physics and astronomy

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

By submitting this form, you accept the Mollom privacy policy.

SAVE PREVIEW