Big Data at the Large Hadron Collider

Frank Würthwein

Professor of Physics
University of California San Diego
May 6th, 2015
The Science of the LHC
The Universe is a strange place!

~68% of energy is “dark energy”
We got no clue what this is.

~27% of “energy” is “dark matter”
We have some ideas but no proof of what this is!

All of what we know makes up
Only about 5% of the universe.
To study Dark Matter we need to create it in the laboratory.
The Large Hadron Collider (LHC)

- 27 km in circumference
- Colliding protons on protons at energies of 7, 8, 13, 14 TeV
- 2808 bunches colliding every 25 ns
- with 115 billion protons per bunch
The CMS Experiment

100 Megapixel “camera” …
… taking 40 Million “pictures” per second
… of which 1/40,000 is kept for offline analysis.

several 10’s of Petabytes of data expected per experiment in next Run (2015-2017).

Collaboration between
180 Institutions from 40 countries
“Big bang” in the laboratory

- We gain insight by colliding particles at the highest energies possible to measure:
  - Production rates
  - Masses & lifetimes
  - Decay rates
- From this we derive the “spectroscopy” as well as the “dynamics” of elementary particles.
- Progress is made by going to higher energies and brighter beams.
LHC Science during the last 5 years

• Analyze the official experiment data (~10PB) to reduce it to custom data (~400TB)
  – bring data we need to UCSD Mayer Hall cluster
    • enough disk space to keep things as long as we felt like it
  – store all our private data at UCSD Mayer Hall as well
  – do analysis of private data at UCSD

• In the next 5 years, data volumes are expected to grow large enough that we need to be more agile.
LHC science in the next 5 years

• Be much more agile !!!
• Cache data temporarily at UCSD for analysis
• Access data via the WAN
  – compute at UCSD on data stored elsewhere
  – compute elsewhere on data stored at UCSD
• Compute at SDSC on data at Mayer Hall
• Need high performance IDI for all of the above
On 4/28 network shown to support 70Gbps

Collaboration between:
- Prism (SDSC/CalIT2)
- Cherub (SDSC/ACT/CENIC)
- ESNet - DOE

![Graph showing network traffic (04/28/15)]
Progress Last Week

On 4/28 network shown to support 70Gbps

On 5/4 Mayer Hall cluster shown to support 40Gbps from Caltech

Traffic in Gbps:
- 70 Gbps
- 60 Gbps
- 50 Gbps
- 40 Gbps
- 30 Gbps
- 20 Gbps
- 10 Gbps

Traffic details:
- Peak inbound traffic: 69.63 Gbps (87%)
- Peak outbound traffic: 71.19 Gbps (89%)
- 100% Bandwidth (80 Gbps)
Progress Last Week

On 4/28 network shown to support 70Gbps

We’re off to a good start!

However, there is a lot more work to be done to integrate the new capabilities into our production operations!!!
Summary & Conclusions

• LHC Science generates 10’s to 100’s of Petabytes of Data within the next 5-10 years

• To be effective, we need to be agile
  – bring data to PB cache at Mayer Hall
  – Compute at SDSC on data in Mayer Hall
  – Compute outside UCSD on data in Mayer Hall
  – Compute in Mayer Hall on data outside UCSD

• To succeed requires national & international collaboration, including strong IDI at UCSD.

Thanks!