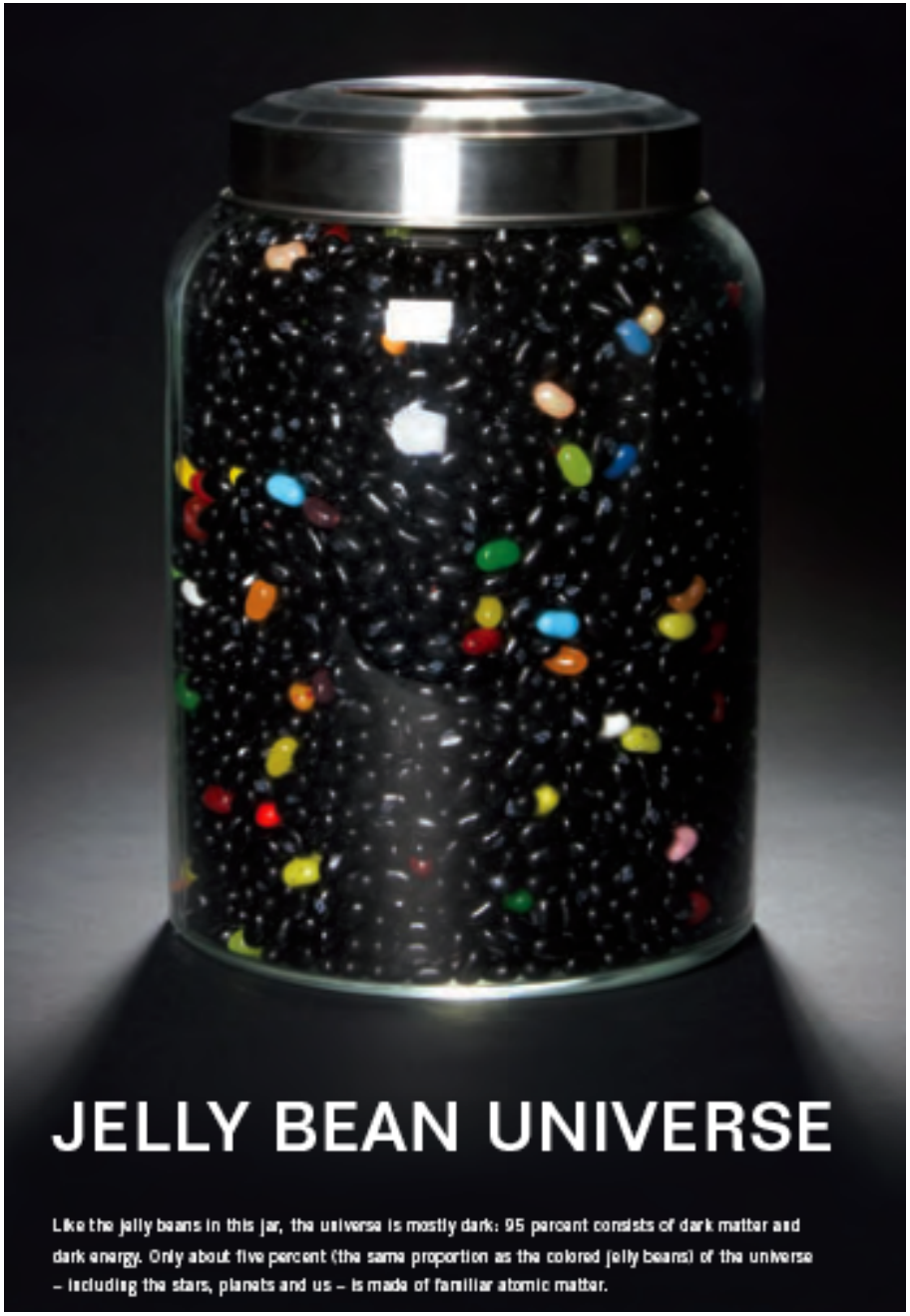
Big Data at the Large Hadron Collider

Frank Würthwein

Professor of Physics
University of California San Diego
May 6th, 2015

The Science of the LHC



The Universe is a strange place!

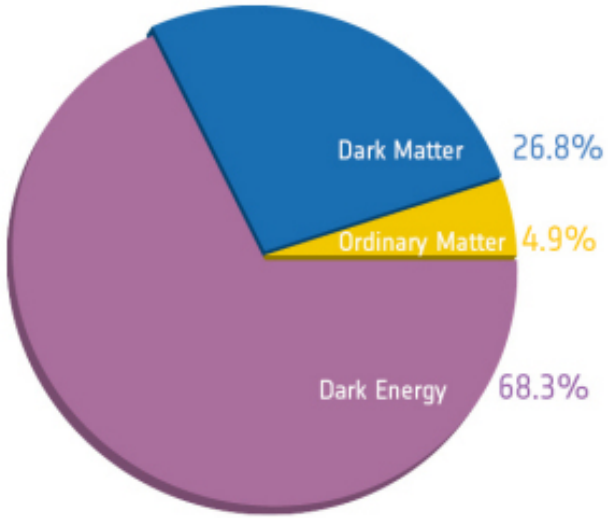
~68% of energy is "dark energy"

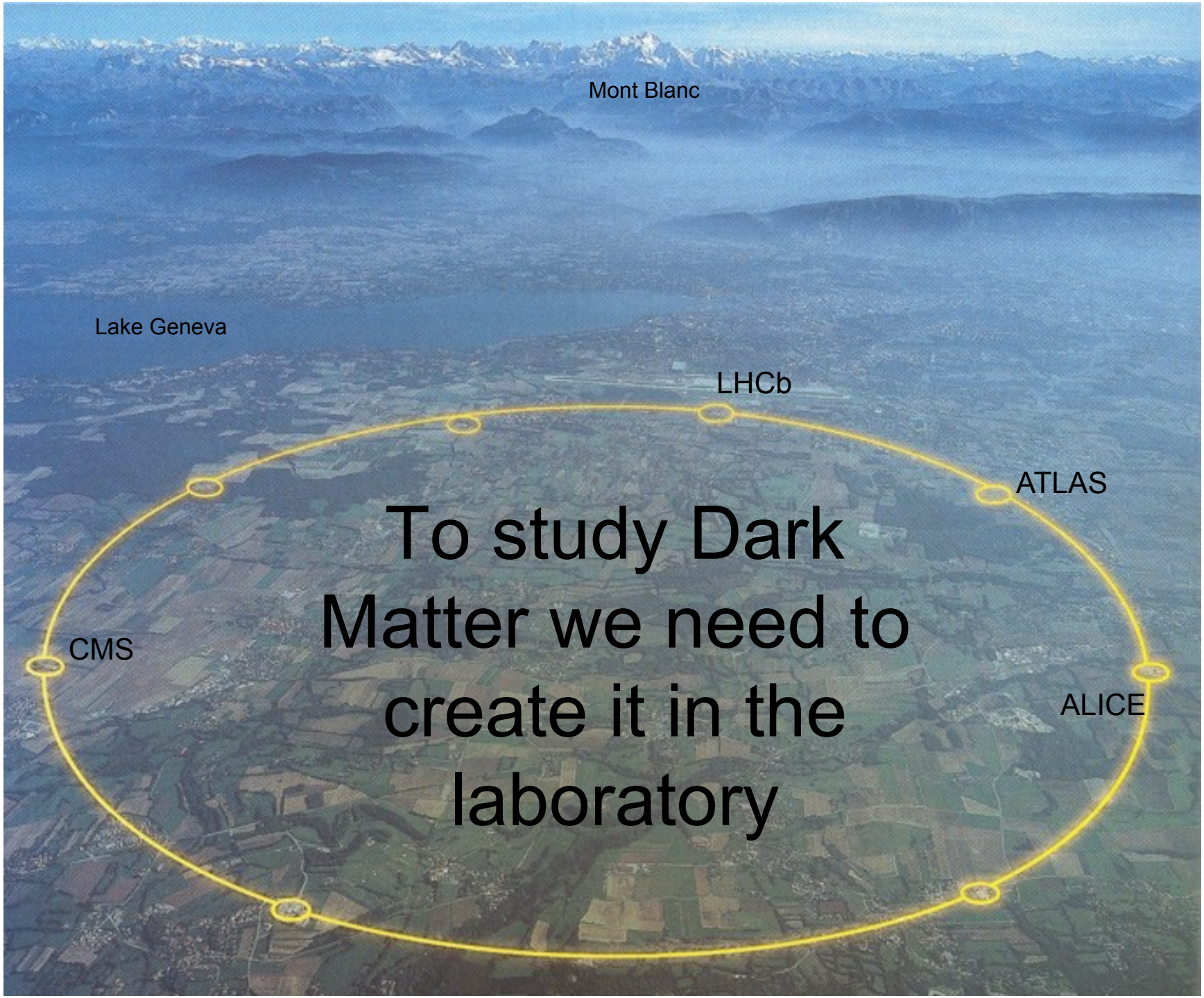
We got no clue what this is.

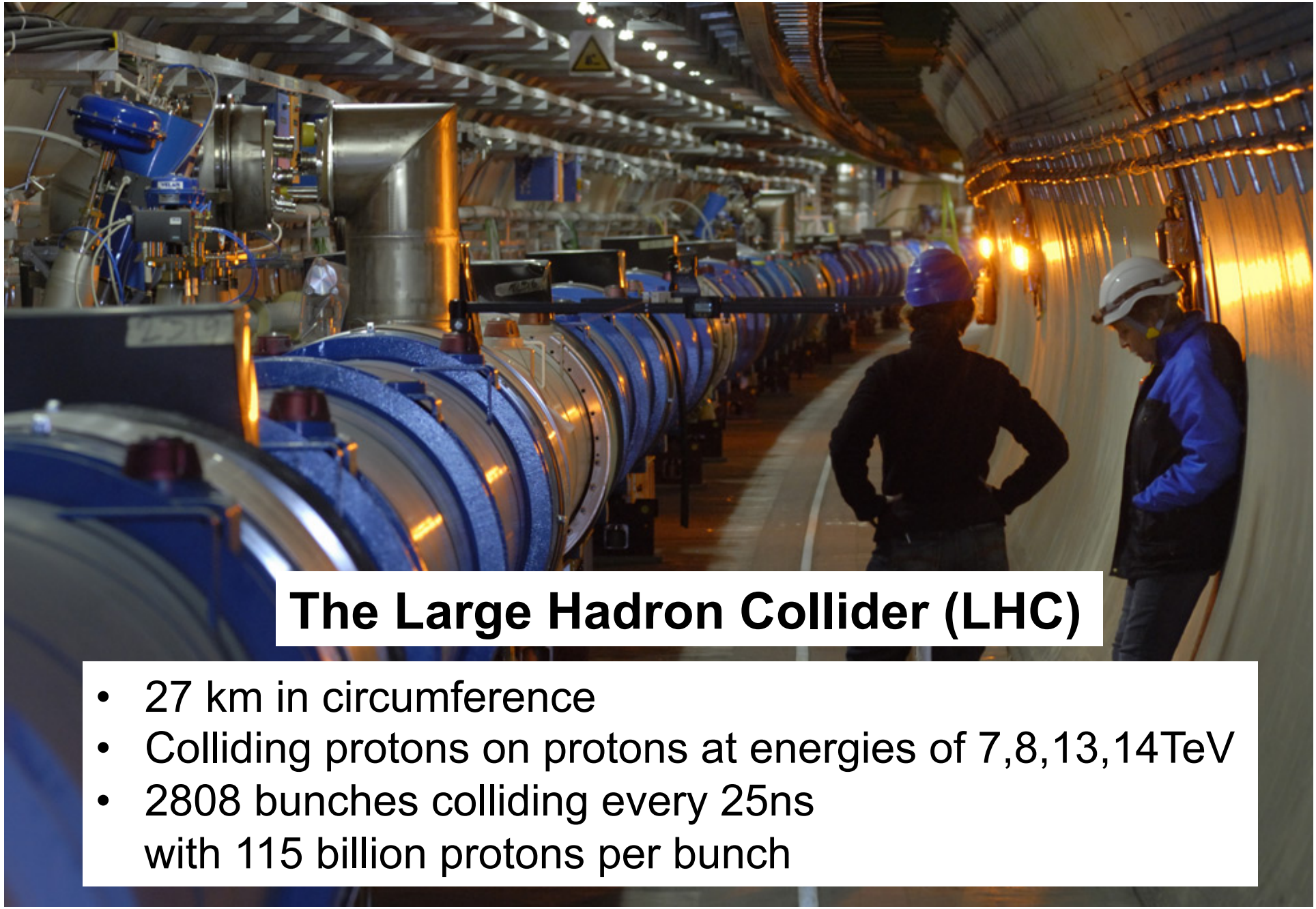
~27% of "energy" is "dark matter"

We have some ideas but no proof of what this is!

All of what we know makes up Only about 5% of the universe.




Mont Blanc
Lake Geneva
LHCb
ATLAS
CMS
ALICE
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 7,8,13,14TeV
- 2808 bunches colliding every 25ns
  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.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

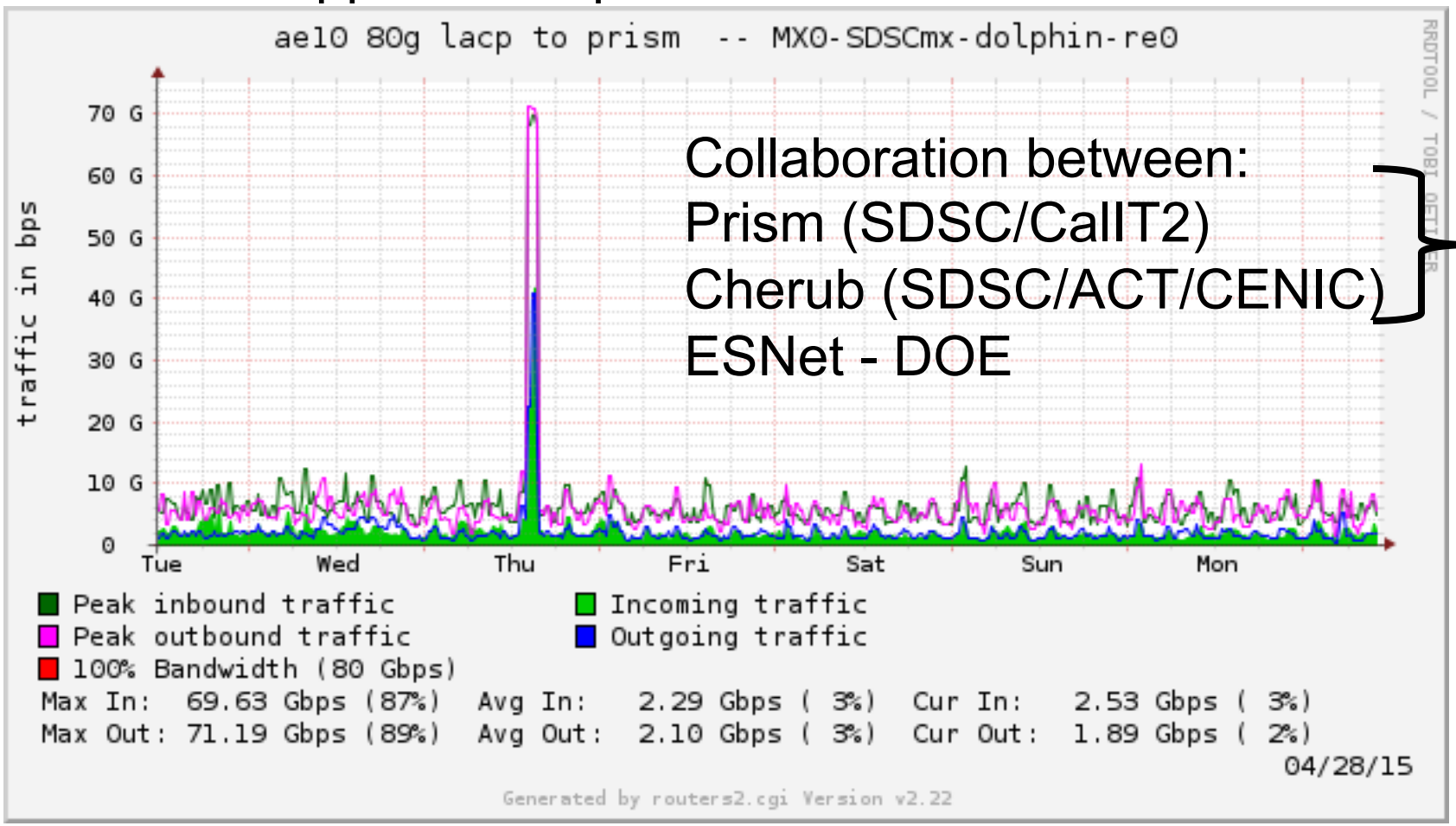
LHC science in the next 5 years

- **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

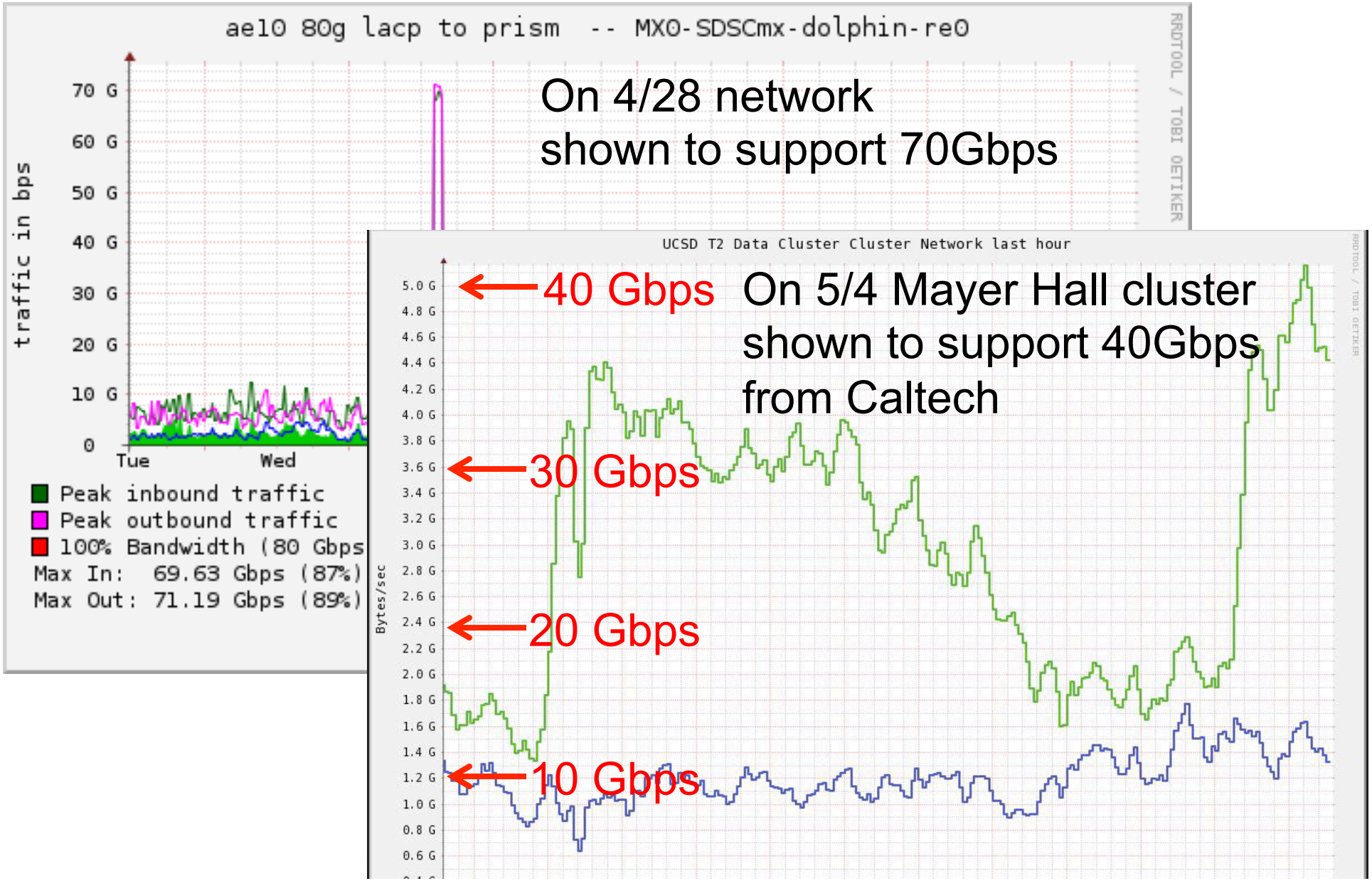
SDSC SAN DIEGO SUPERCOMPUTER CENTER

Progress Last Week

On 4/28 network
shown to support 70Gbps



Progress Last Week



Progress Last Week

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 !

SDSC SAN DIEGO SUPERCOMPUTER CENTER