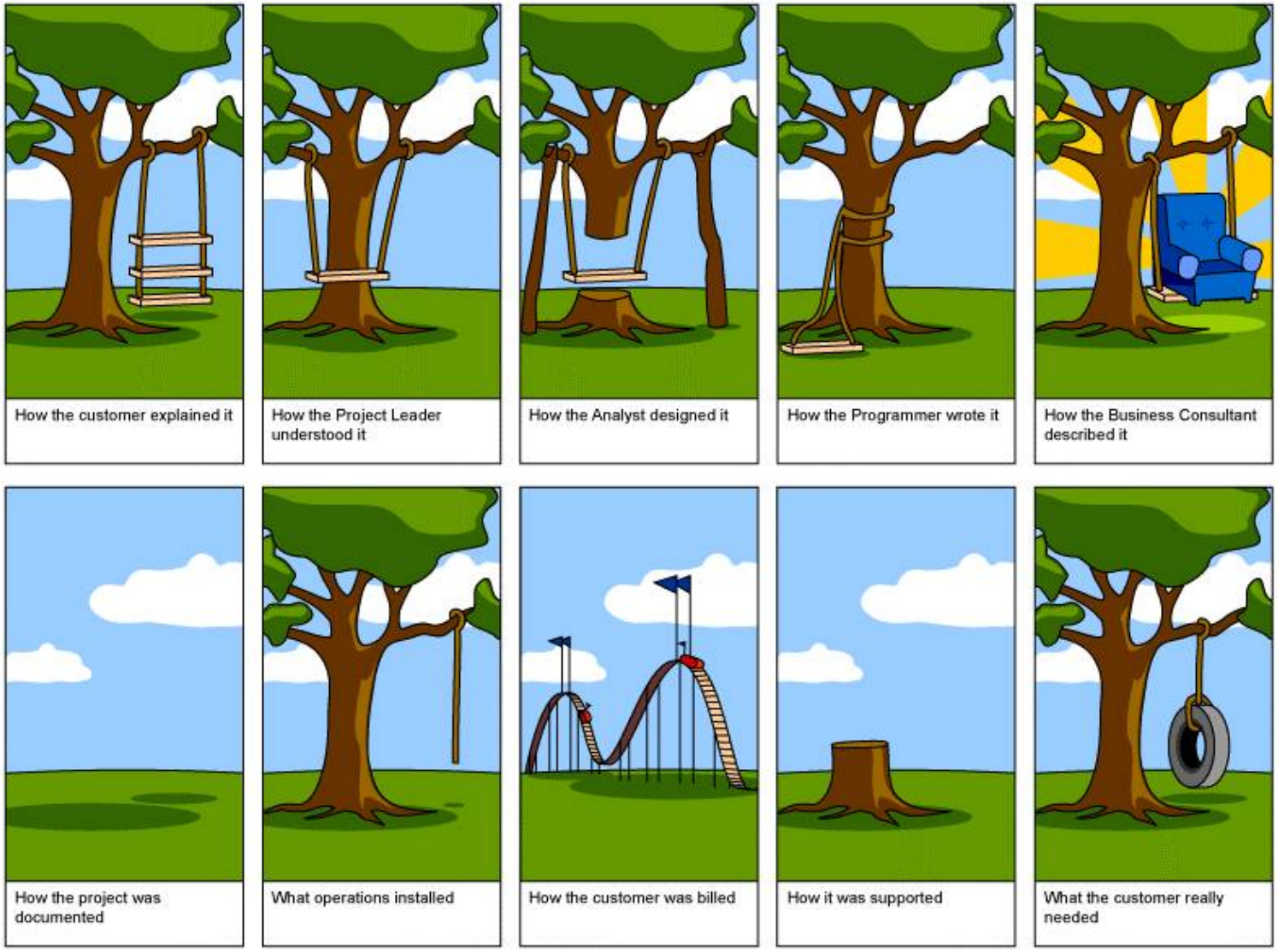
High performance End2End:

Derivations from first principles & Demo proposal for LHCONE

Inder Monga
Chief Technologist & Area Lead
Energy Sciences Network

Date: May 3rd, 2013

Objective of working together: Bridging the impedance mismatch, Between Applications & Networks



Classic impedance mismatch

CARES ABOUT

Throughput

CAN ONLY PROVIDE

Bandwidth

CANNOT ASK THE NETWORK TO PROVIDE A HIGH-THROUGHPUT TRANSFER

How do we provide the right network capabilities that make it easier for the application to get better throughput?

Getting Better requires effort

– Are we happy with best effort?

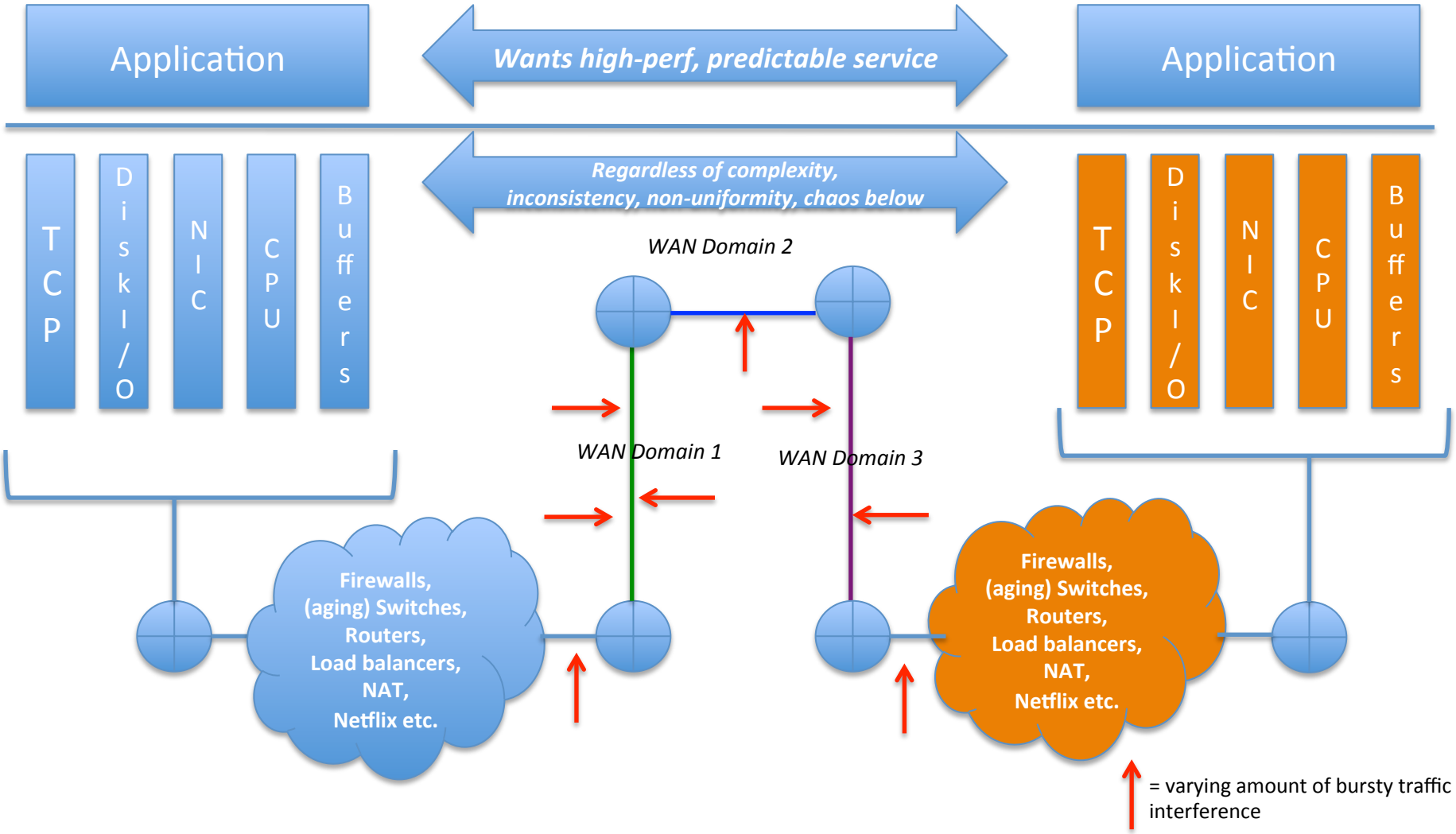
OR

– Do we desire 'better than best' effort?

1 TB data transfer SHOULD take:
10 Mbps network : 300 hrs (12.5 days)
100 Mbps network : 30 hrs
1 Gbps network : 3 hrs
10 Gbps network : 20 minutes

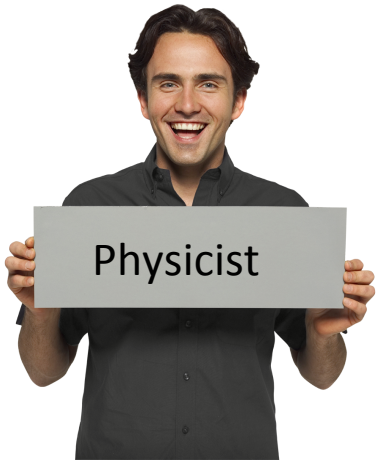
- Networks have not been an issue for LHC so far because people desired 'better than best'

(Seriously) Defining the problem

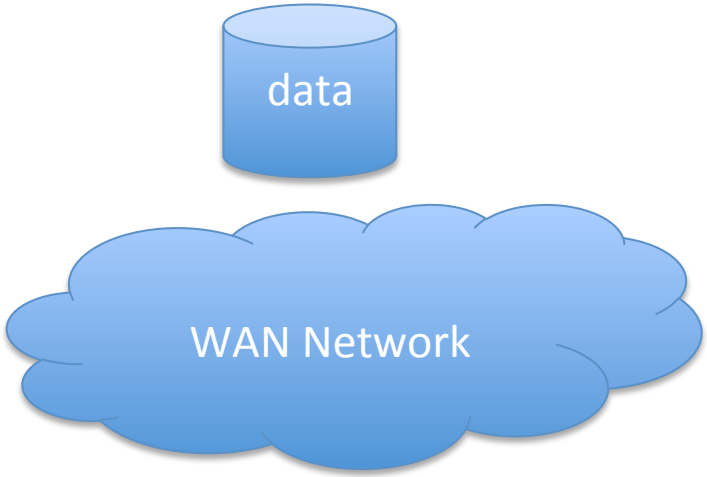


This is an end-to-end, complex systems management problem

& expectations management as well



I found the Higgs!

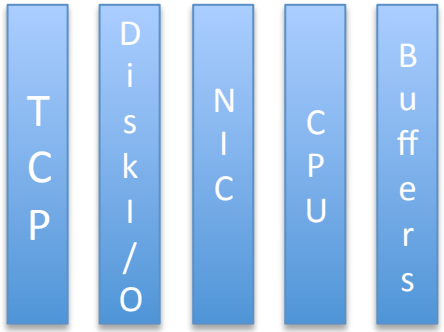


Compute/ Analysis

Missing Campus network folks

Approach (1/3)

1. Buy right hardware and tune your end hosts



- Data Transfer Node architecture/design/Tuning

  Example:
  http://fasterdata.es.net/science-dmz/DTN/reference-implementation/

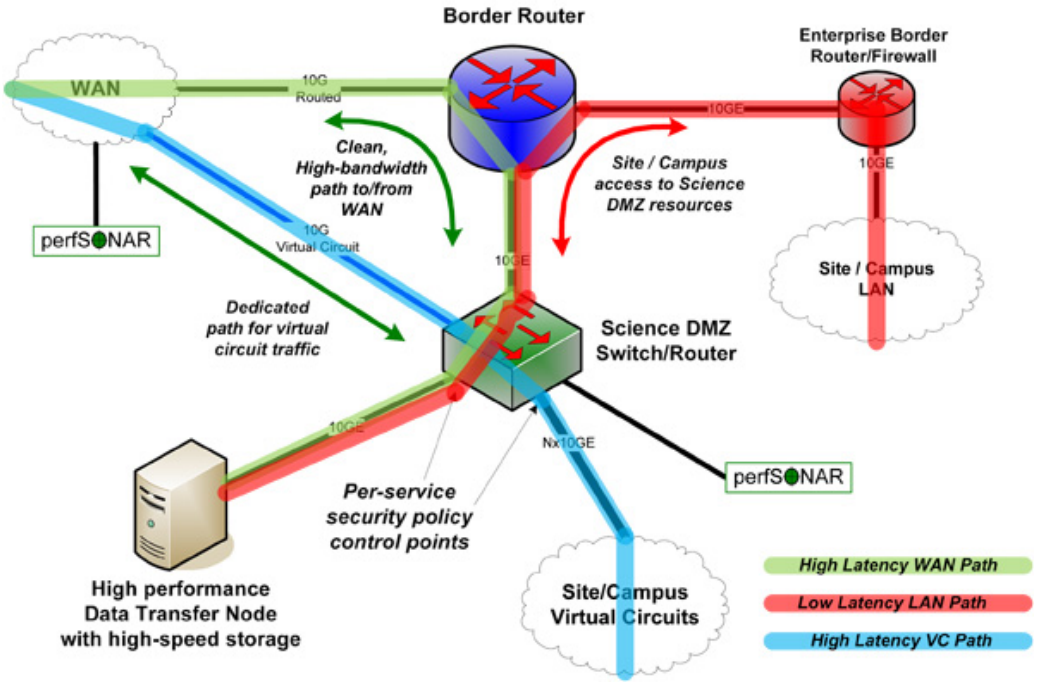
- Choose the right Data Transfer tool

  You are the experts!

  TCP need to be tuned for high RTT, just the best, clean pipes won't give you the greatest throughput

Approach (2/3)

2. Reduce the variability on campus
   - **Science DMZ**: place equipment close to border, eliminate firewall, campus traffic, campus equipment dependence



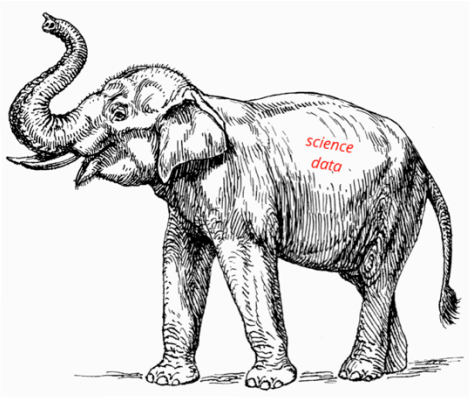
Approach (3/3)

3. Build a lossless Wide Area Path

Effect of 0.0046% packet loss (1 out of 22000 packets) on data transfer rates for *elephant* and *mouse* flows.[1]

How to build a lossless network?

- ample network capacity
- carefully-chosen infrastructure
- deep packet buffers
- automatic and continual verification of network health
- **'fast lanes'**



80x reduction in data transfer rate at DOE-relevant distances (ANL to NERSC) and speeds (10Gpbs).

Negligible.

- **The main goal of LHCONE P2P workshop:**
  - Consistent WAN guaranteed service across multiple NREN domains globally
- Solving only a piece of the overall puzzle, cannot measure benefit without implementing the other pieces

What does NSI provide?

- Consistent guaranteed bandwidth across multiple Wide Area Network domains
  - **Does not guarantee** consistent throughput
  - **Provides a path** different from shortest path (traffic engineering)
- Dynamic is not a requirement
  - Adaptability is
    - How does the network adapt to changing application requirements?
    - CPU, Storage locations change
    - Network capacity is different

Demo proposal for LHCONE

- Choose a few sites that have folks with the desire to experiment with P2P circuits
- Build a static mesh of P2P circuits between the sites with close to zero bandwidth
- Use NSI 2.0 mechanisms to
  - Dynamically increase and reduce bandwidth
  - Based on Job placement or transfer queue to that particular site
  - Based on dynamic allocation of resources

Measuring Success

- How will PhEDEx (for example) know if P2P helped improve anything?
  - Closed feedback loop is important
  - Calibration of what's possible is important
  - Deploy perfSONAR hosts to measure one way latency, and active bandwidth tests
  - Compare application throughput to best possibility