BDII Story
An Evolutionary Approach

Laurence Field
CERN

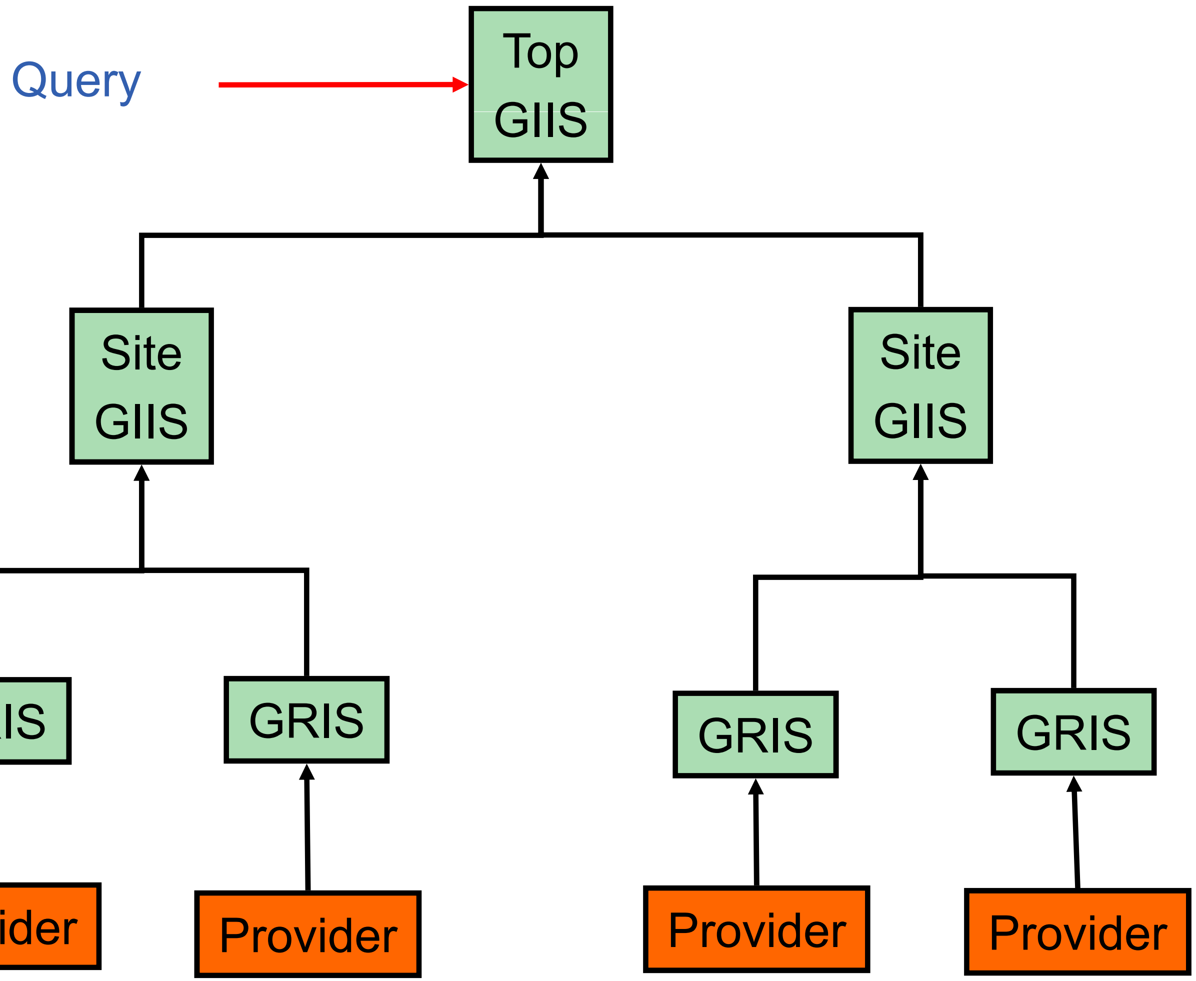
Contents

- **MDS Experience**
- **BDII Evolution**
- **Development Tips**
  - Fundamentals
  - Key Messages

Globus MDS v2

- **Metadata Directory Service (MDS)**
  - http://www.globus.org/toolkit/docs/2.4/mds/
- **Information Providers (IP)**
  - Scripts that get the information and return LDIF
- **Grid Resource Information Service (GRIS)**
  - Daemon that runs the IP and answers LDAP queries
  - Register to a GIIS
- **Grid Information Index Service (GIIS)**
  - Answers LDAP queries by querying registered GRIS's or GIIS's.
- **Both the GRIS and GIIS have a 30s cache**
  - To reduce load and improve performance

Original MDS Deployment



MDS Problems

- **Query forwarded to lower levels if cache is stale**
  - Complex timings required
    - Which were initial broken
    - Strange behaviour due to timing interaction
  - Lower level problems affected higher levels
    - 1 broken work node could bring down the whole system
      - *qstat hung, IP hung, GRIS hung, GIIS hung*
  - Response significantly slower if cache is not used
    - Seriously affected query scalability
- **MDS did not work in a distributed environment**
    - Initially would not scale past 4 sites
- **OpenLDAP to the rescue!**
  - Tests showed that a standard LDAP server was more stable

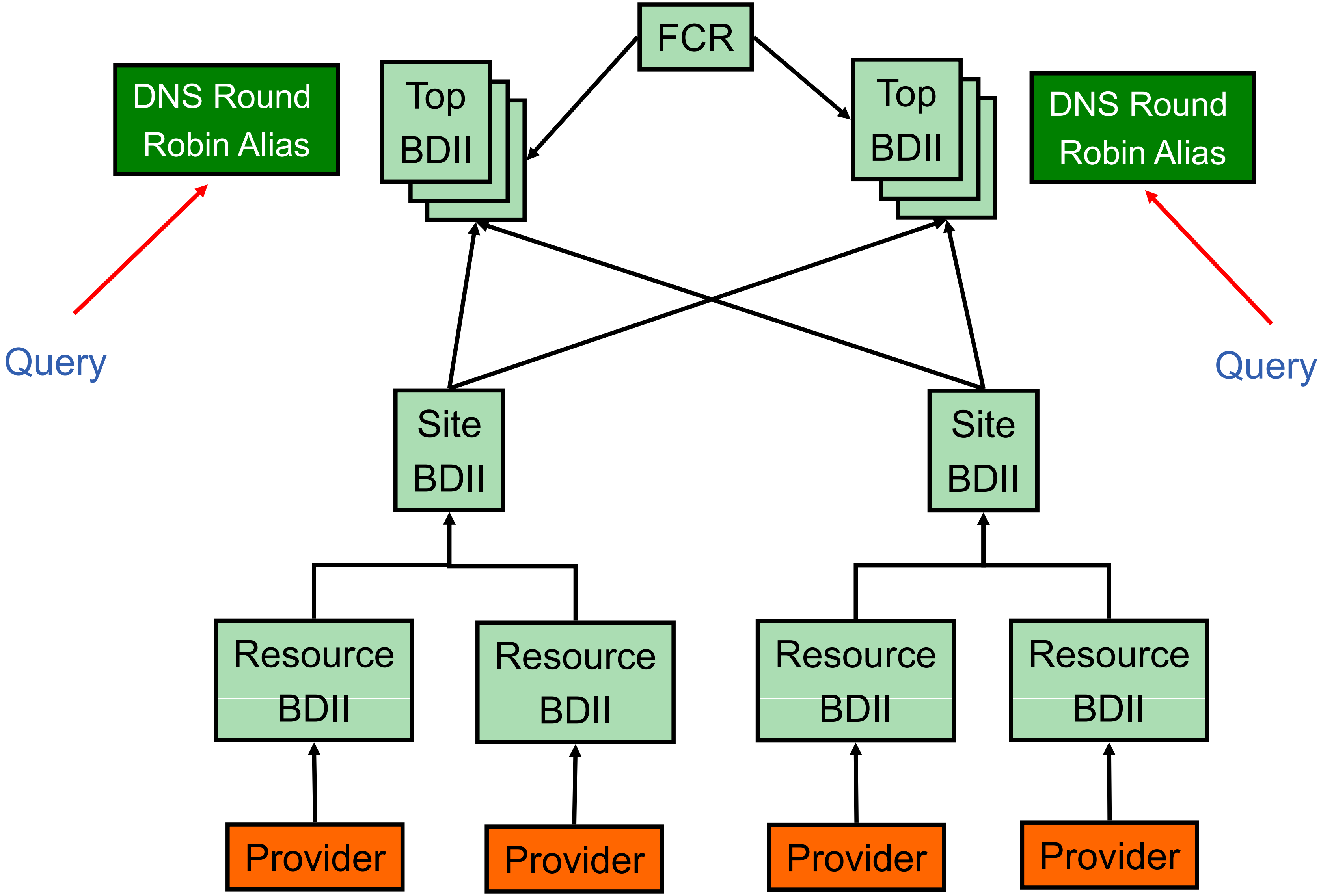
Introducing The BDII

1. **Need to decouple query handling and data gathering**
2. **Use off-the-shelf, established components**

- **The Berkeley Database Information Index**
  - Uses a standard OpenLDAP server as supplied by the OS
    - With the Berkeley database backend
  - Updated by a perl process
    - Configuration file containing LDAP URLs for the sites
    - Use ldapsearch command used as it is stable
      - *And should not break with and OS upgrade*
  - Balance freshness of information and performance

- **BDII first used as top-level GIIS**
  - Due to instability problems of the top-level GIIS
    - Is now used at the top, site and resource level

Information System Architecture



BDII Improvements

- **Initial data gathering done in series**
  - Worked with 5 stable sites
  - However, Each “timed-out” site would added 30s.
    - Firewalled ports are very common!
  - Performance testing done using parallel threads.

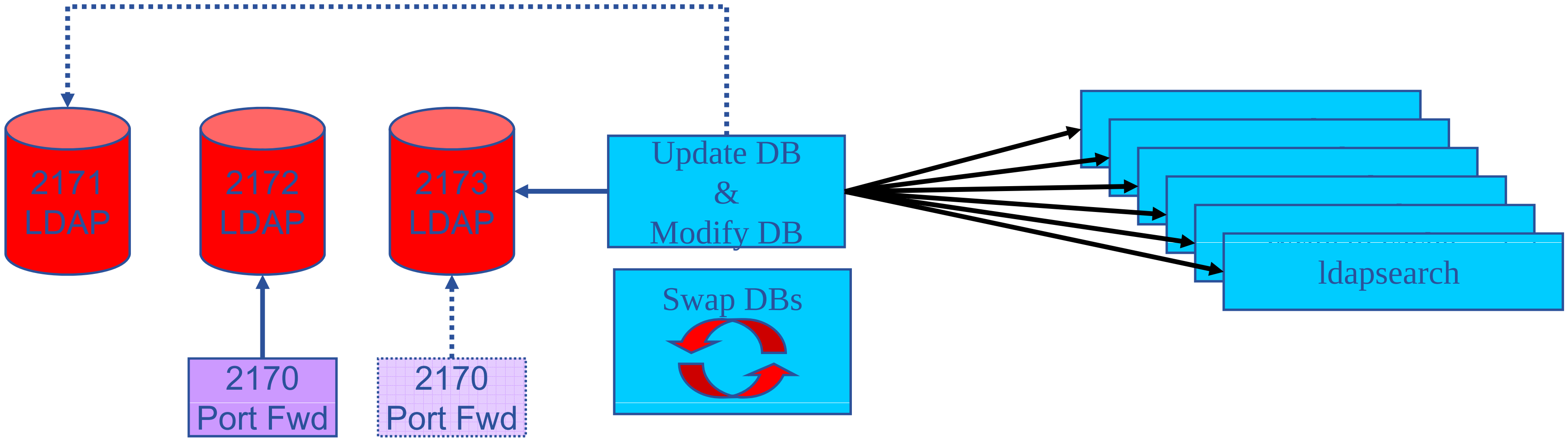
3. Use parallel thread pool to address timeout failures

LDAP ADD scalability testing

  - http://lfield.web.cern.ch/lfield/public/papers/Chep2004.pdf
  - ldapadd faster than ldapmodify
    - drop database and recreate
  - slapadd faster than ldapadd
    - Fastest with bulk updates

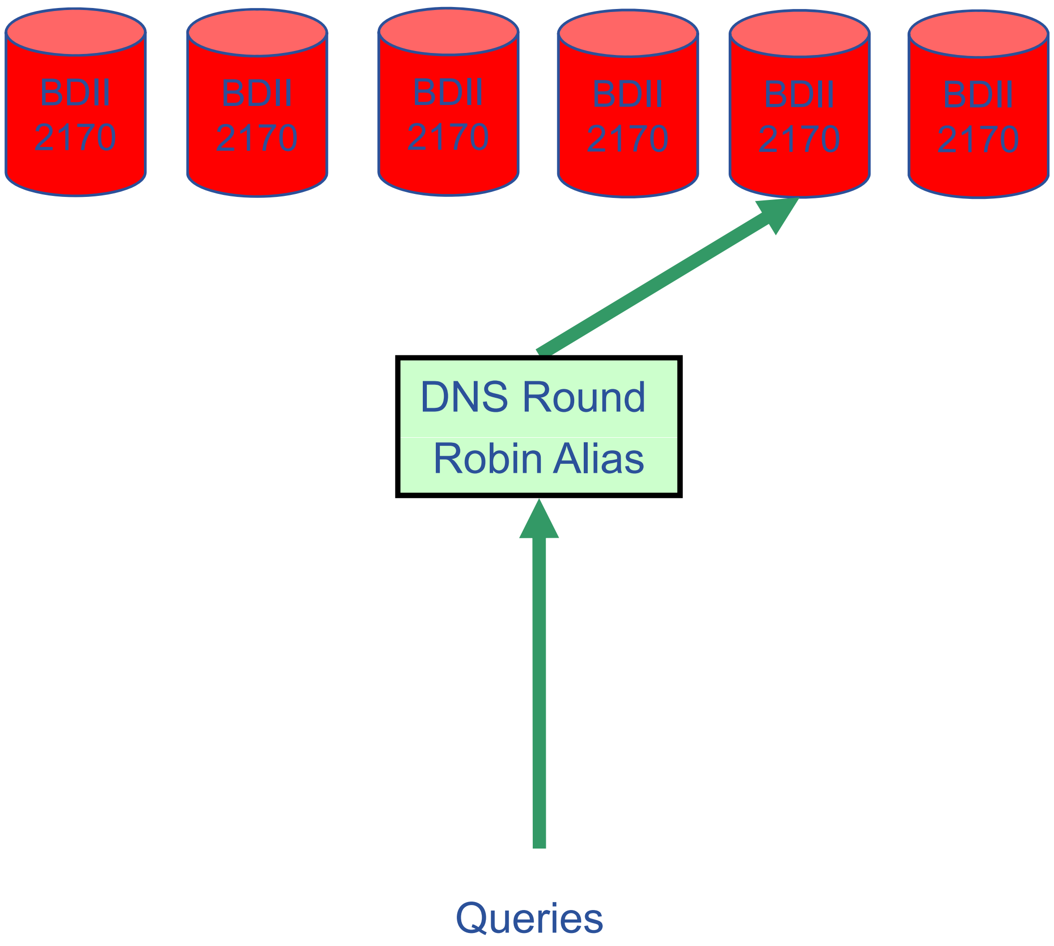
3. Use most performant method, not the most elegant

The BDII Architecture



- **Multiple DBs instances used to increase performance**
  - Writing to the DB affects the read operations
  - Separate read and write
  - Read only, write only and one spare for queries to finish
  - This functionality is enabled by the port forwarder
- **List of LDAP sources to query from configuration file**

Load Balancing The BDII

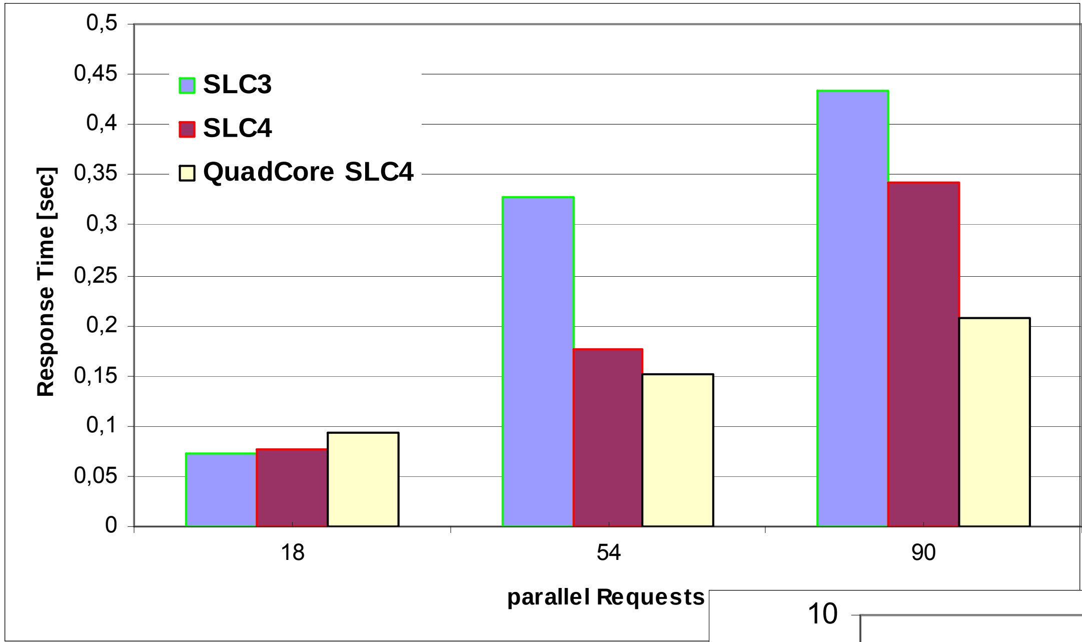


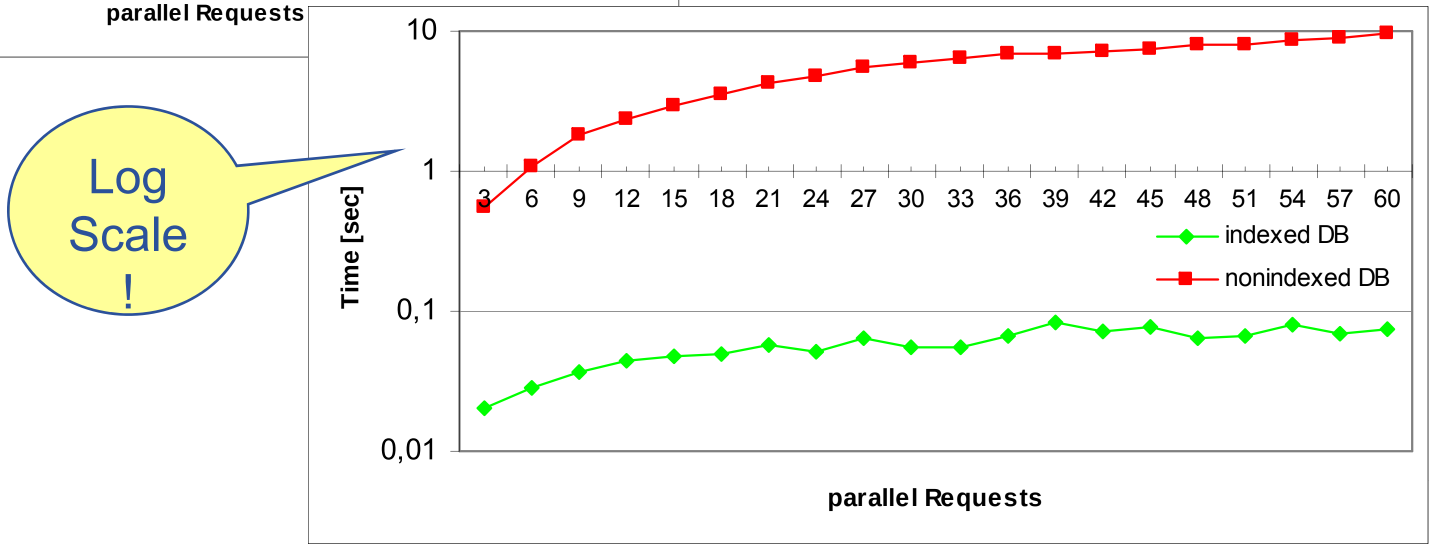
Query Scalability Improvements

- **Log queries to production BDIIs**
  - Analyzed log files for production usage
  - Improved client queries
- **Designed query performance tests**
  - Using production queries as the input
- **Evaluated different deployment scenarios**
  - http://lfield.web.cern.ch/lfield/public/papers/CHEP2007-Performance.pdf
    - Hardware
    - OpenLDAP Versions
    - Slapd configuration
      - *Using indexes significantly improved performance*

5. **Understand and measure the real usage**

Performance Improvements





Future Work

- **Ongoing technology evaluation**
  - Is OpenLDAP the most performant technology?
- **Improving performance through service deployment**
  - Improve load balancing mechanism
    - Alternative handling of state

6. **Separate state and connection handling**

- **Re-evaluate the architecture**
  - Can we improve the performance through design?
    - Based on the real usage measured use cases

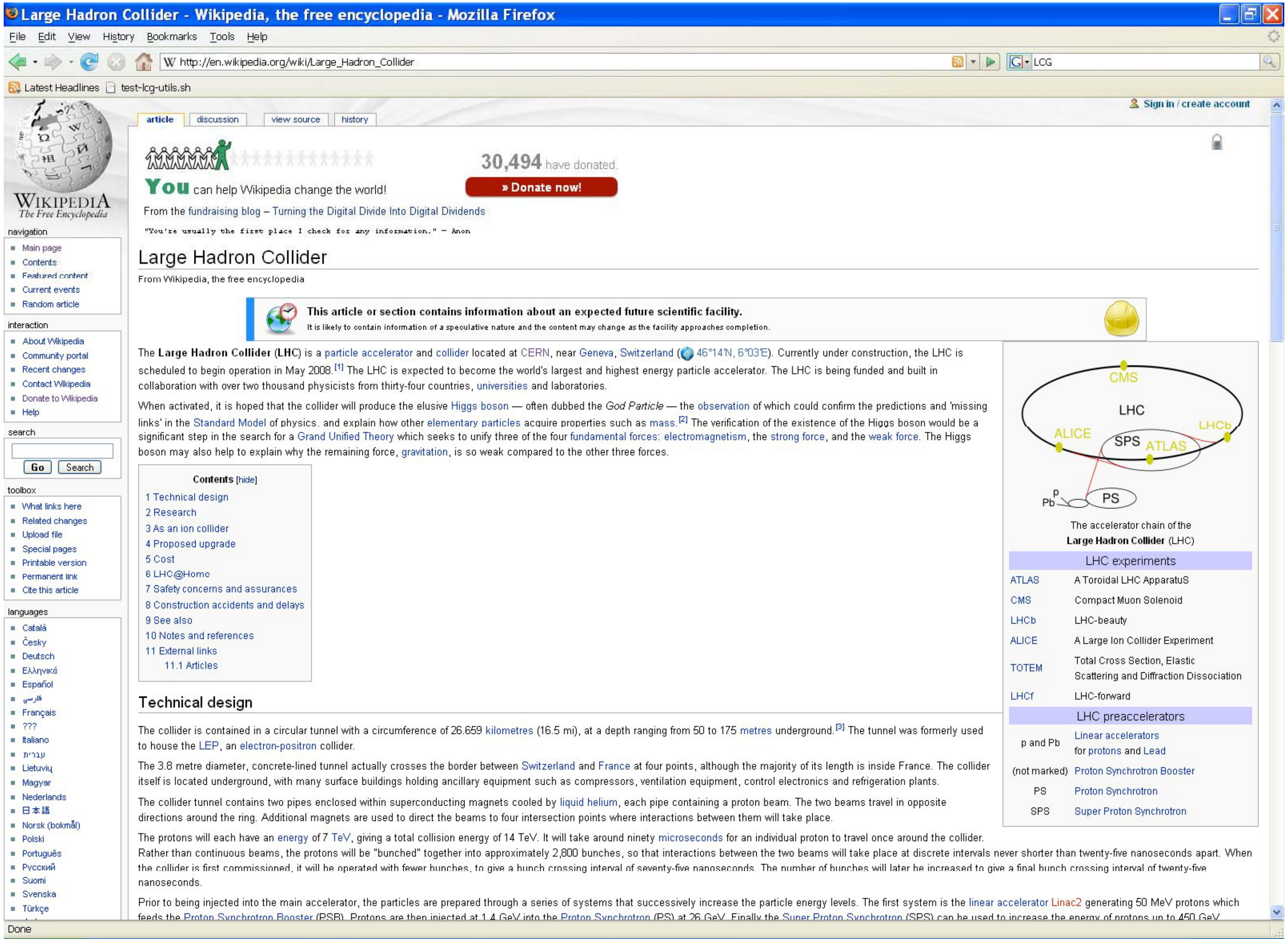
1. **Scalability**
   - Over subscribed service tend to be brittle

2. **Robustness**
   – $P_{failure}(x) \rightarrow 1$ as $x \rightarrow \infty$

Key Messages

- **Keep It Simple Stupid (KISS)**
  - Complexity increases the chance of failure
- **Focus on the core task**
  - Make it robust make it scale
    - Avoid any feature enhancements until the core is robust
  - Understand and measure the production use cases
- **Build upon established technology/standards**
  - Avoid emerging technology/standards
    - and unstable/unproven software
  - Use the appropriate technology
    - XML is not the only solution!
  - We are never the first to solve basic computing problems!
- **Understand the management of connections and state**
  - Well managed state is easy to load balance
    - The different between software and a service

Software



Large Hadron Collider - Wikipedia, the free encyclopedia - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://en.wikipedia.org/wiki/Large_Hadron_Collider

article | discussion | view source | history

30,494 have donated.

You can help Wikipedia change the world!

» Donate now!

From the fundraising blog – Turning the Digital Divide Into Digital Dividends

"You're usually the first place I check for any information." – Anon

Large Hadron Collider

From Wikipedia, the free encyclopedia

This article or section contains information about an expected future scientific facility.
It is likely to contain information of a speculative nature and the content may change as the facility approaches completion.

The **Large Hadron Collider** (**LHC**) is a particle accelerator and collider located at CERN, near Geneva, Switzerland (46°14′N, 6°03′E). Currently under construction, the LHC is scheduled to begin operation in May 2008.[1] The LHC is expected to become the world's largest and highest energy particle accelerator. The LHC is being funded and built in collaboration with over two thousand physicists from thirty-four countries, universities and laboratories.

When activated, it is hoped that the collider will produce the elusive Higgs boson — often dubbed the *God Particle* — the observation of which could confirm the predictions and 'missing links' in the Standard Model of physics, and explain how other elementary particles acquire properties such as mass.[2] The verification of the existence of the Higgs boson would be a significant step in the search for a Grand Unified Theory which seeks to unify three of the four fundamental forces: electromagnetism, the strong force, and the weak force. The Higgs boson may also help to explain why the remaining force, gravitation, is so weak compared to the other three forces.

Contents [hide]

The accelerator chain of the **Large Hadron Collider** (LHC)

LHC experiments	
ATLAS	A Toroidal LHC ApparatuS
CMS	Compact Muon Solenoid
LHCb	LHC-beauty
ALICE	A Large Ion Collider Experiment
TOTEM	Total Cross Section, Elastic Scattering and Diffraction Dissociation
LHCf	LHC-forward
LHC preaccelerators	
p and Pb	Linear accelerators for protons and Lead
(not marked)	Proton Synchrotron Booster
PS	Proton Synchrotron
SPS	Super Proton Synchrotron

Technical design

The collider is contained in a circular tunnel with a circumference of 26.659 kilometres (16.5 mi), at a depth ranging from 50 to 175 metres underground.[3] The tunnel was formerly used to house the LEP, an electron-positron collider.

The 3.8 metre diameter, concrete-lined tunnel actually crosses the border between Switzerland and France at four points, although the majority of its length is inside France. The collider itself is located underground, with many surface buildings holding ancillary equipment such as compressors, ventilation equipment, control electronics and refrigeration plants.

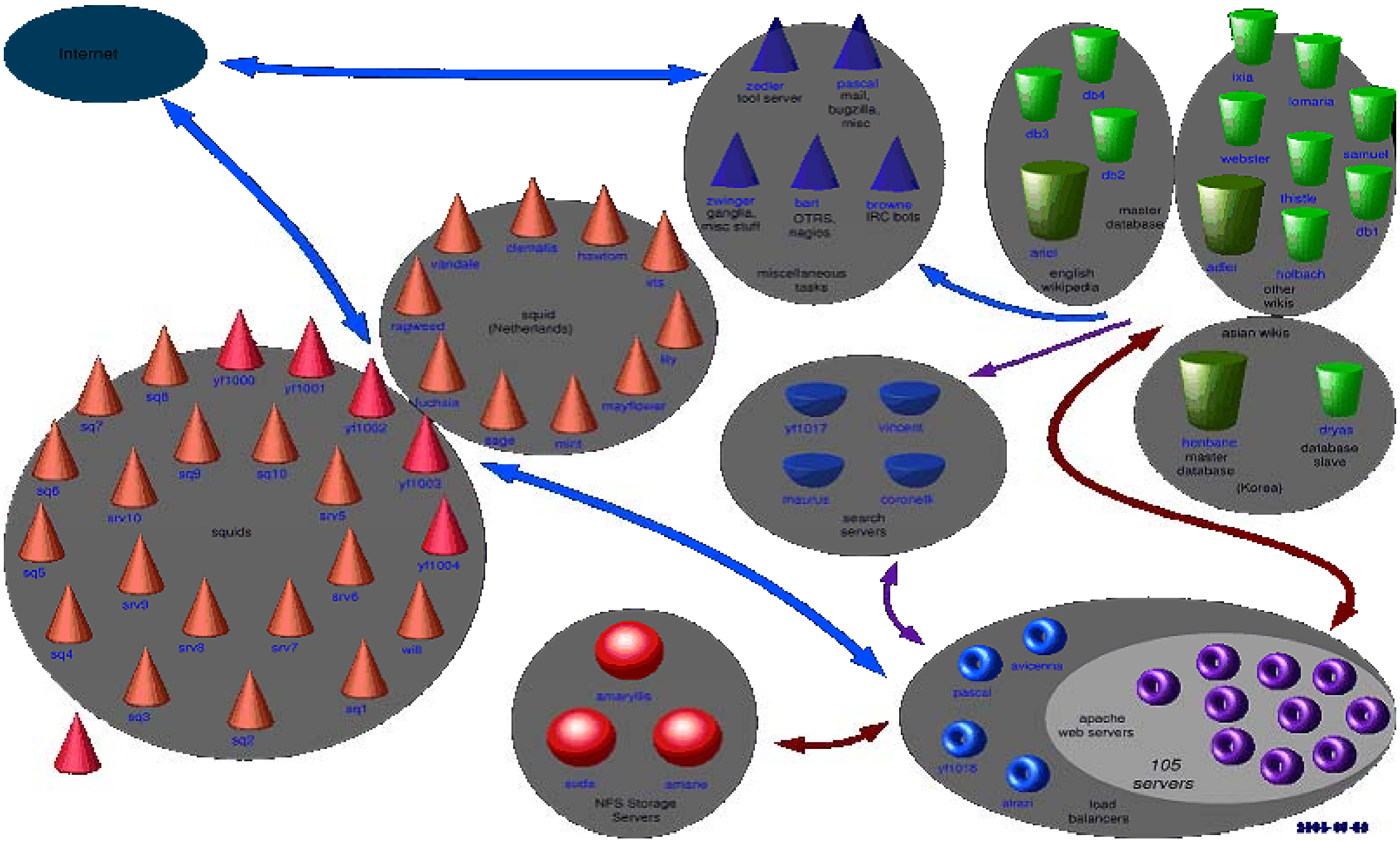
The collider tunnel contains two pipes enclosed within superconducting magnets cooled by liquid helium, each pipe containing a proton beam. The two beams travel in opposite directions around the ring. Additional magnets are used to direct the beams to four intersection points where interactions between them will take place.

The protons will each have an energy of 7 TeV, giving a total collision energy of 14 TeV. It will take around ninety microseconds for an individual proton to travel once around the collider. Rather than continuous beams, the protons will be "bunched" together into approximately 2,800 bunches, so that interactions between the two beams will take place at discrete intervals never shorter than twenty-five nanoseconds apart. When the collider is first commissioned, it will be operated with fewer bunches, to give a bunch crossing interval of seventy-five nanoseconds. The number of bunches will later be increased to give a final bunch crossing interval of twenty-five nanoseconds.

Prior to being injected into the main accelerator, the particles are prepared through a series of systems that successively increase the particle energy levels. The first system is the linear accelerator Linac2 generating 50 MeV protons which feeds the Proton Synchrotron Booster (PSB). Protons are then injected at 1.4 GeV into the Proton Synchrotron (PS) at 26 GeV. Finally the Super Proton Synchrotron (SPS) can be used to increase the energy of protons up to 450 GeV.

Done

Service


Internet
zedler
tool server
pascal
mail,
bugzilla,
misc
zwinger
ganglia,
misc stuff
bart
OTRS,
nagios
browne
IRC bots
miscellaneous
tasks
db4
db3
db2
ariel
master
database
english
wikipedia
ixia
lomaria
webster
samuel
thistle
db1
adler
holbach
other
wikis
asian wikis
henbane
master
database
dryas
database
slave
(Korea)
vandale
clematis
hawtorn
iris
ragweed
squid
(Netherlands)
lily
fuchsia
mayflower
sage
sq8
yf1000
yf1001
sq7
yf1002
sq6
sq9
sq10
yf1003
srv10
srv5
squids
sq5
yf1004
srv9
srv6
sq4
srv8
srv7
will
sq3
sq1
sq2
yf1017
vincent
maurus
coronelli
search
servers
amaryllis
suda
amane
NFS Storage
Servers
avicenna
pascal
apache
web servers
105
servers
yf1018
alrazi
load
balancers