Showing Real Big Data

Towards visualisation of data transfers with 'Big Data' analytics techniques

HvA induction session February 2016

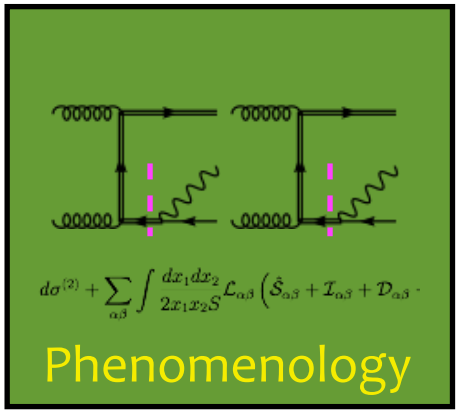
David Groep, Nikhef

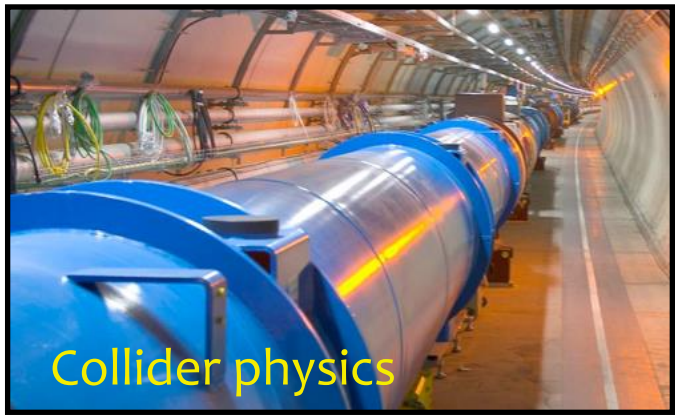
David Groep
Nikhef
Amsterdam
PDP programme

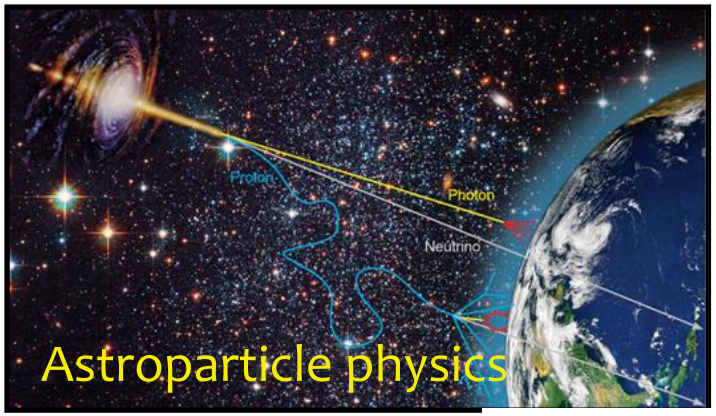
Nationaal instituut voor subatomaire fysica

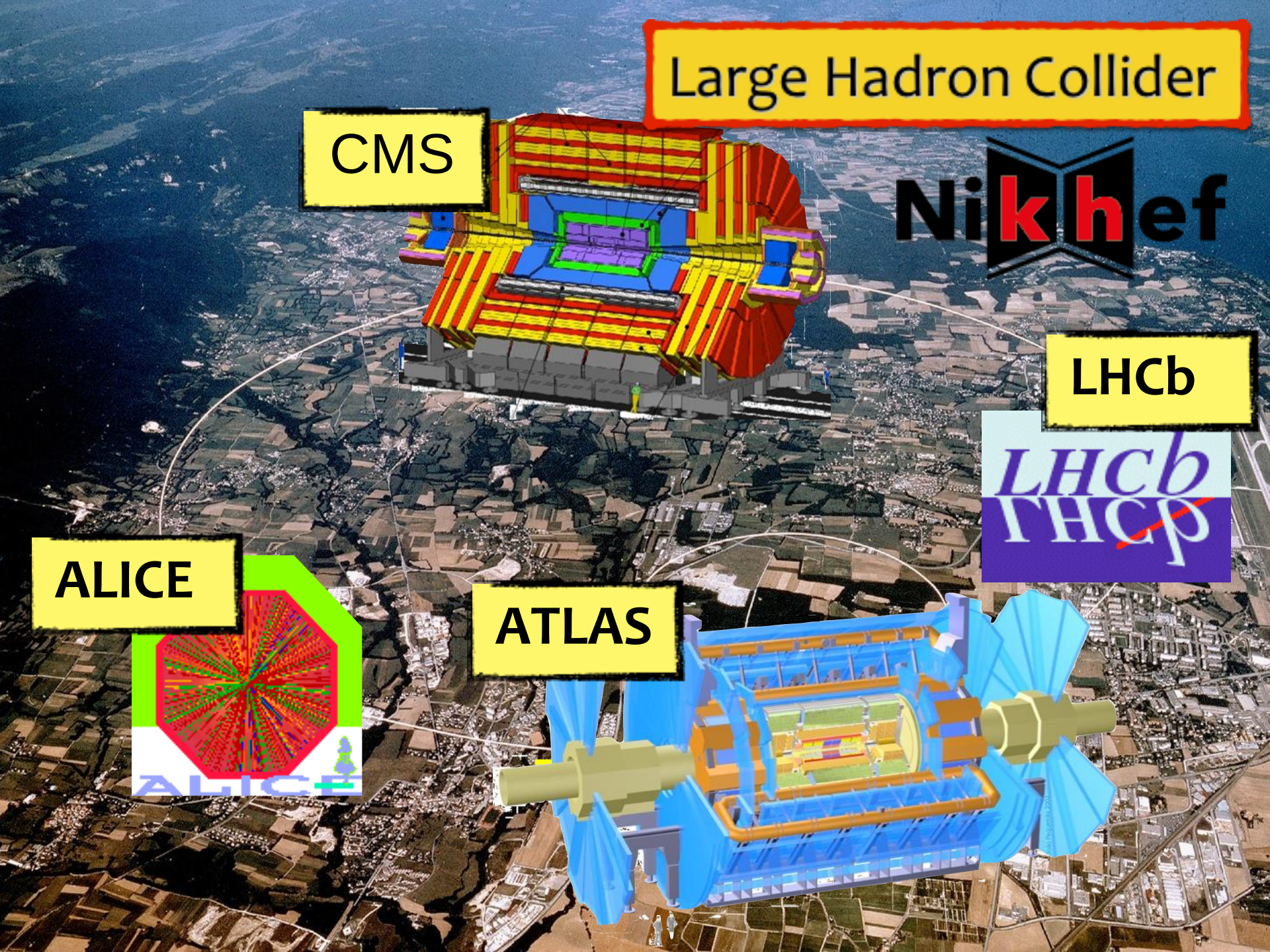
Verleggen van de grenzen van onze kennis

- **Accelerator-based particle physics**
  Experiments studying interactions in particle collision processes at particle accelerators, in particular at CERN;
- **Astroparticle physics**
  Experiments studying interactions of particles and radiation emanating from the Universe.

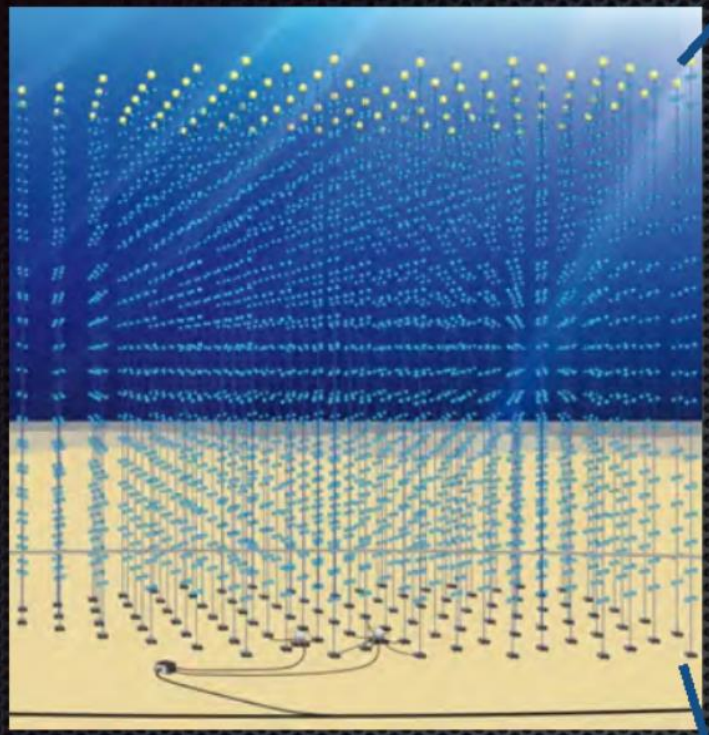


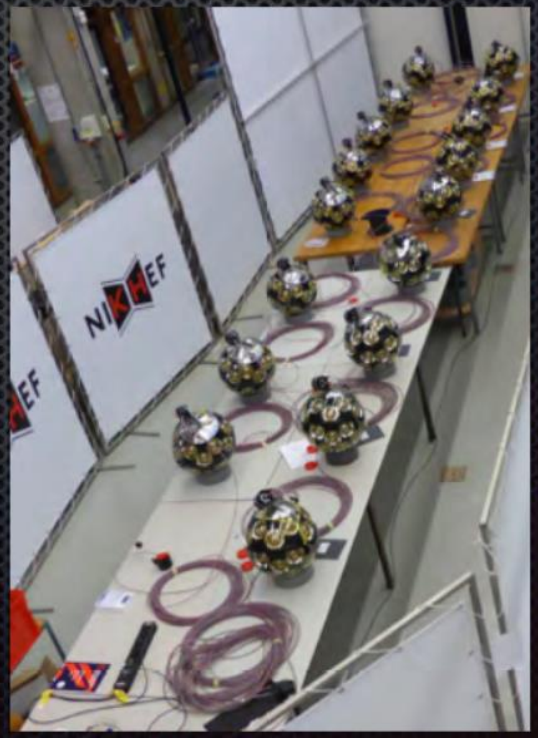





Large Hadron Collider
CMS
Nikhef
LHCb
LHCb
ALICE
ATLAS

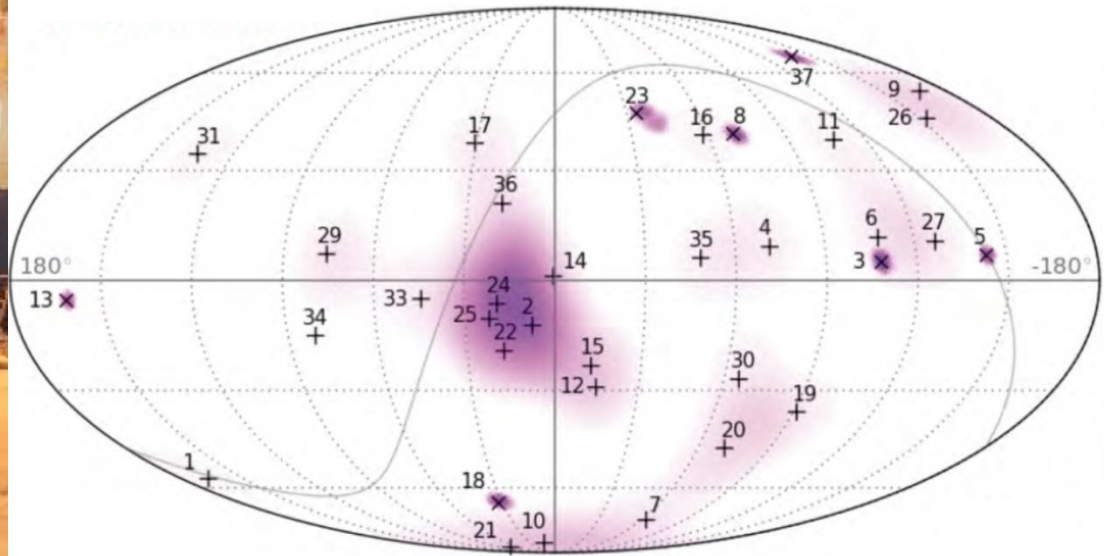
Nikhefs neutrino-detector: KM3NeT

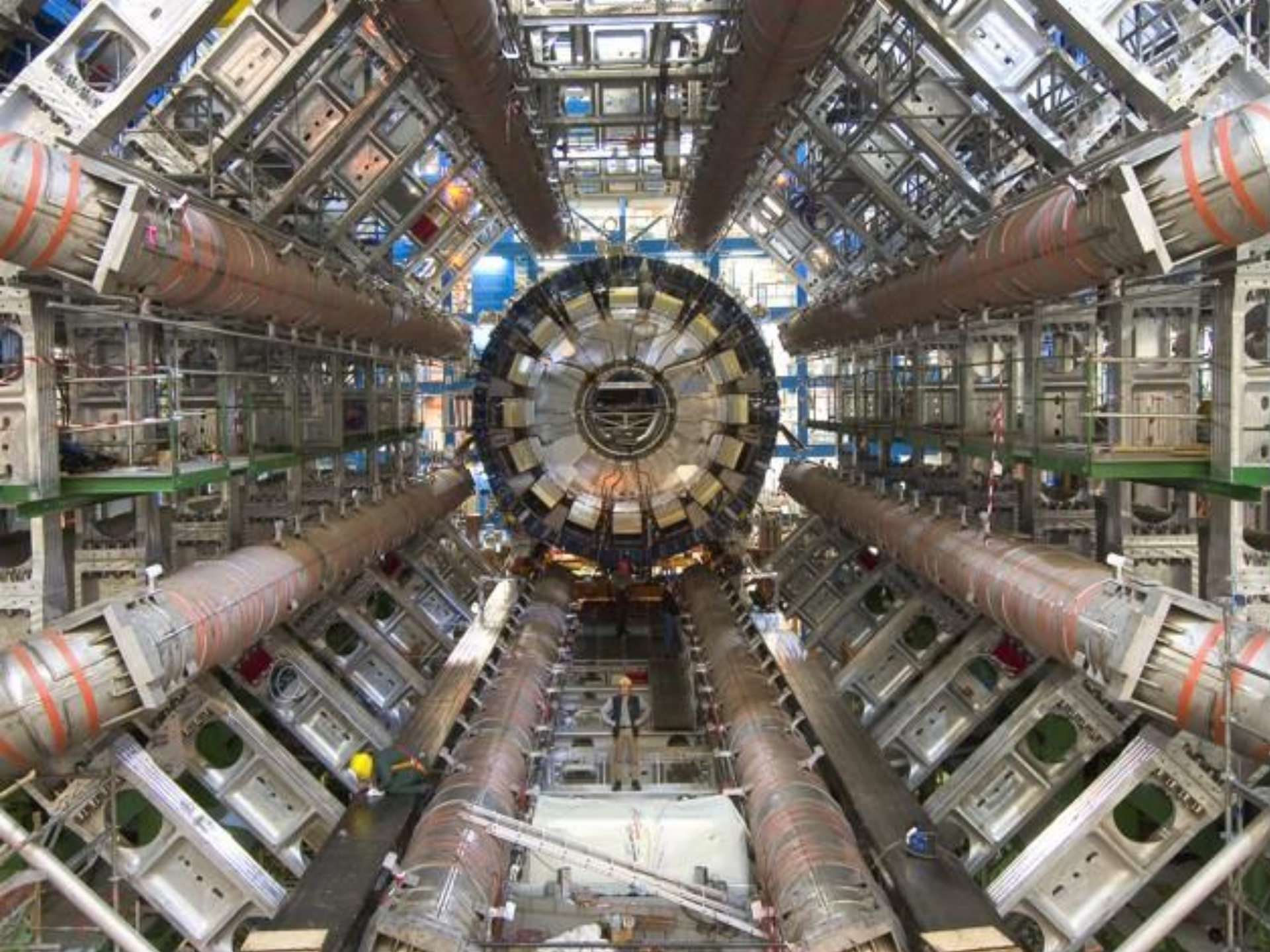




Little white structures prevent the HV bases and cables to touch each other

De Melkweg

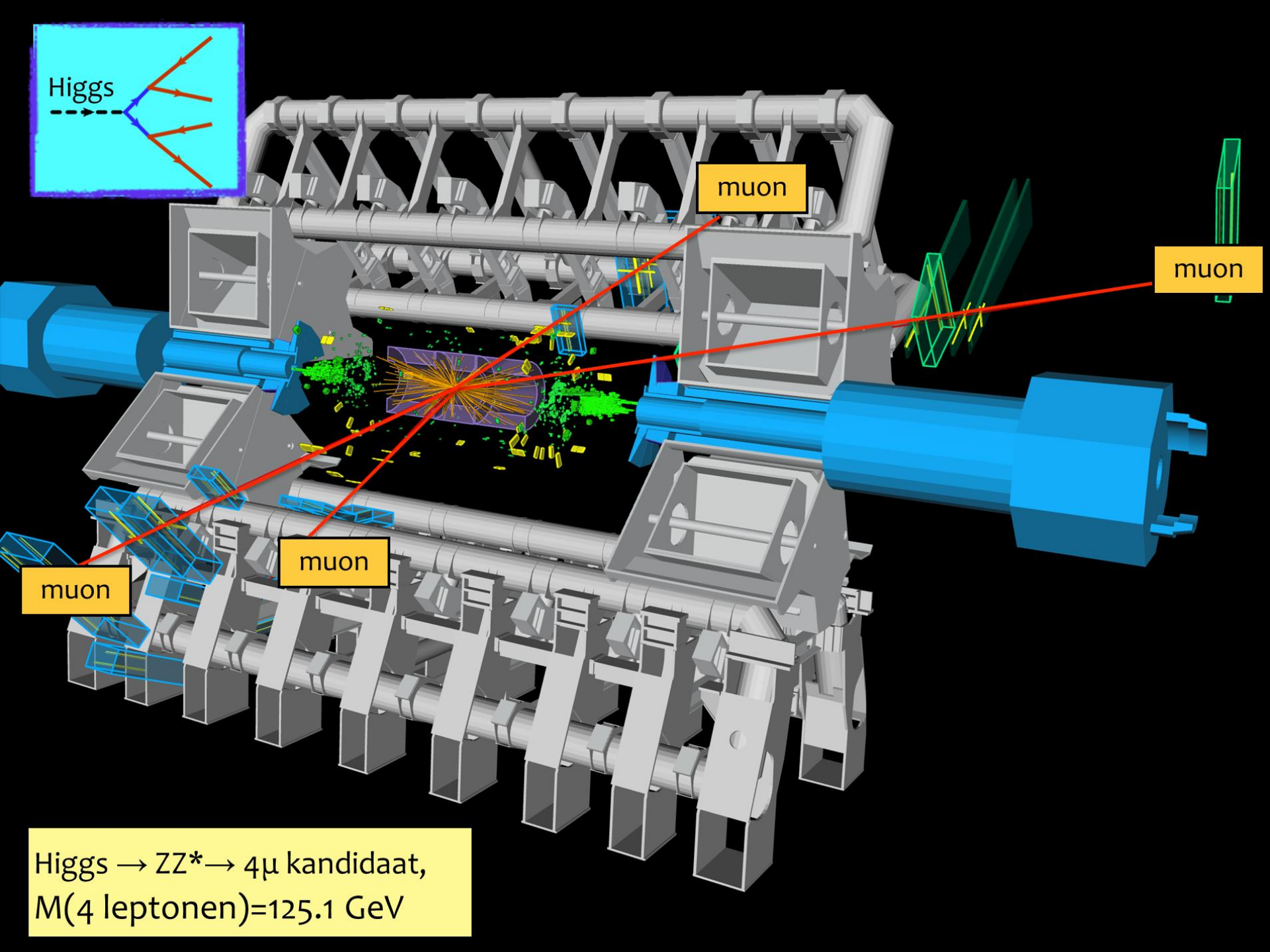




Kans Higgs deeltje:

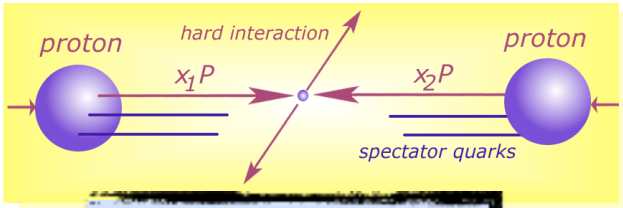
1 op de 1.000.000.000.000 bostingen

- Dit is equivalent met zoeken van 1 persoon op 1000 wereldpopulaties
- Oftewel één naald in 20 miljoen hooibergen

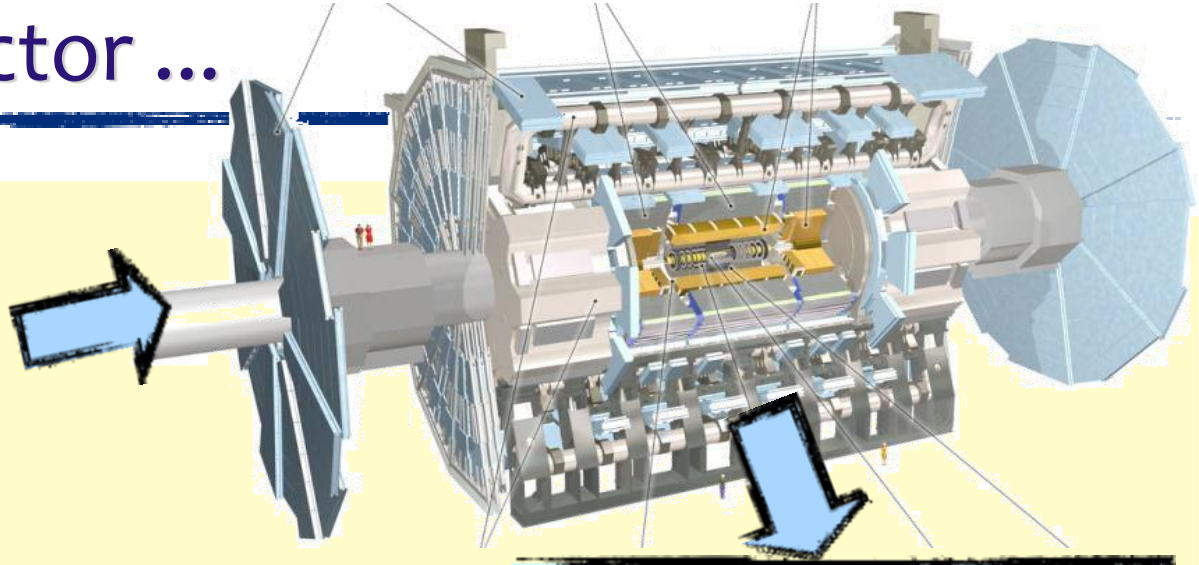


Higgs → ZZ*→ 4μ kandidaat,
M(4 leptonen)=125.1 GeV

Detector to doctor ...

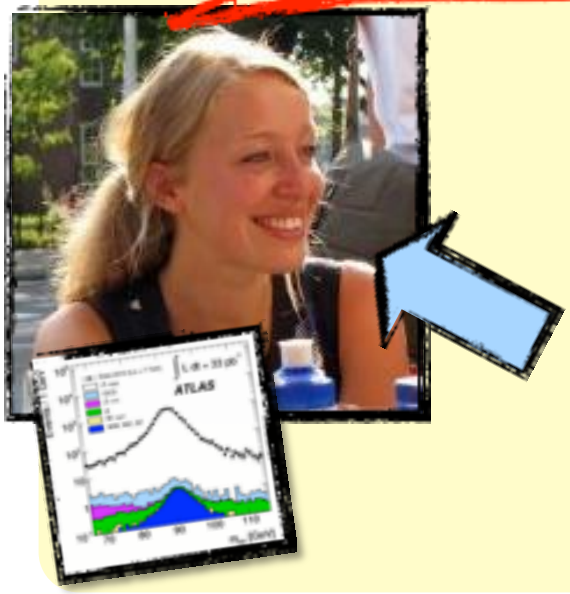


40 miljoen / seconde



Trigger systeem selecteert 600 Hz
~ 1 GB/s data

Analyse van botsingen door promovendi



and processing

Data distributie met GRID computers

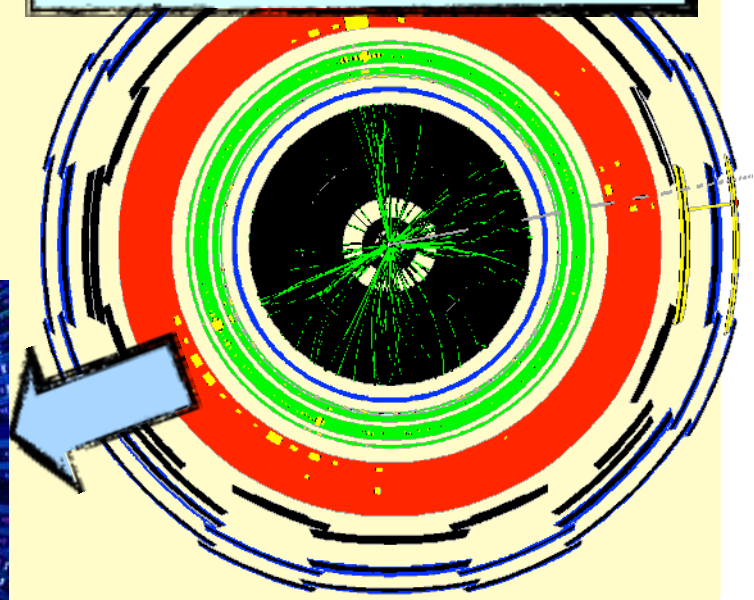


Image source: joint CERN (wLCG) and EGI

Organisations participating in the global collaboration of e-Infrastructures

Even just for wLCG, supporting the CERN LHC programme

More than 200 independent institutes with end-users
More than 50 countries & regions
More than 300 service centres
Handful regional 'service coordination organisations'
300 000 CPU cores, 200+PByte storage
One independent 'policy-bridge' PKI

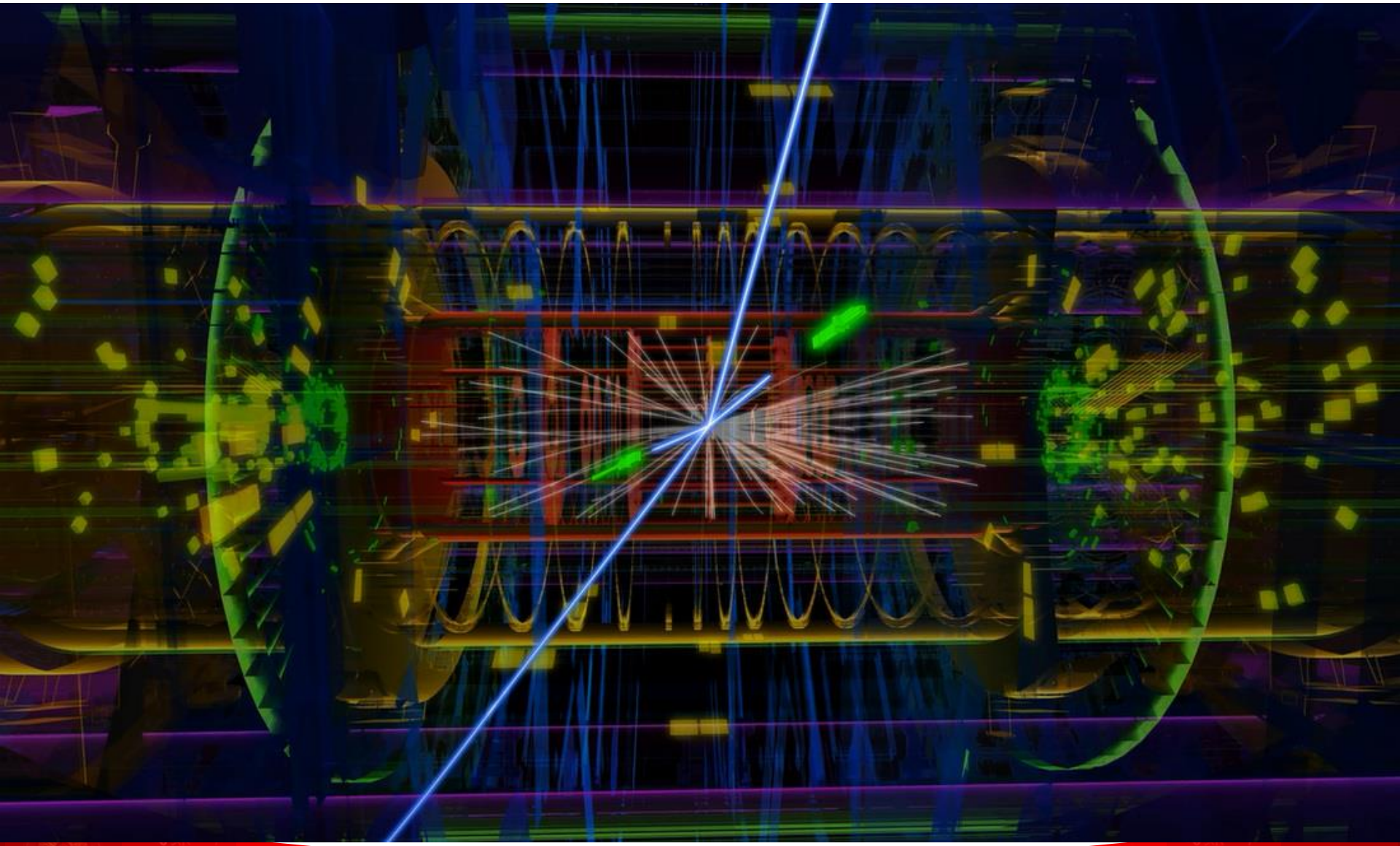
PRACE

Open Science Grid

EGI

XSEDE
Extreme Science and Engineering Discovery Environment

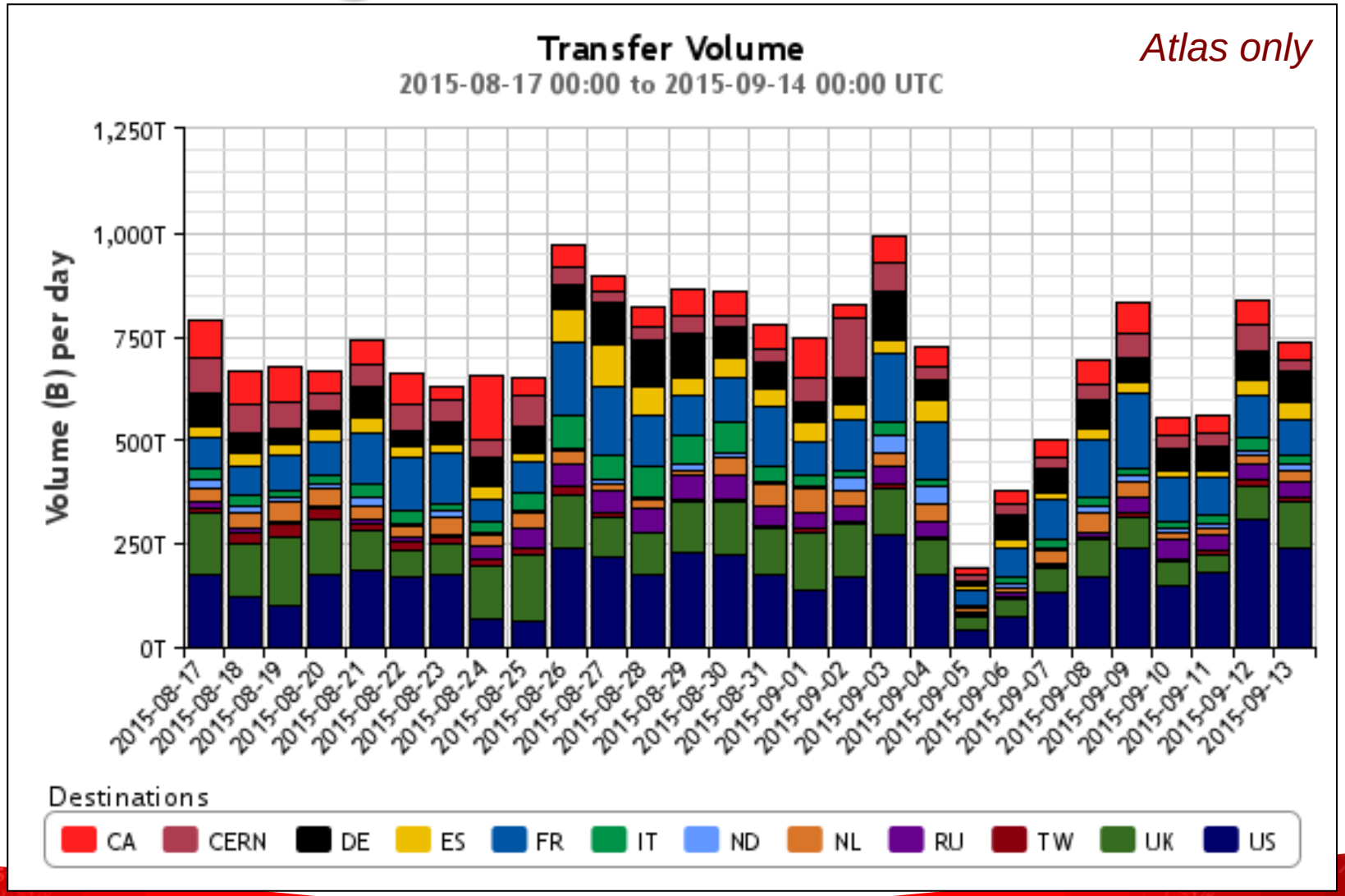
Atlas: ~50 TByte/day raw data to tape; 1000 TByte/day processed data transfers



Nikhef

Image source: CERN, Atlas

Big ‘as in Large’ Data

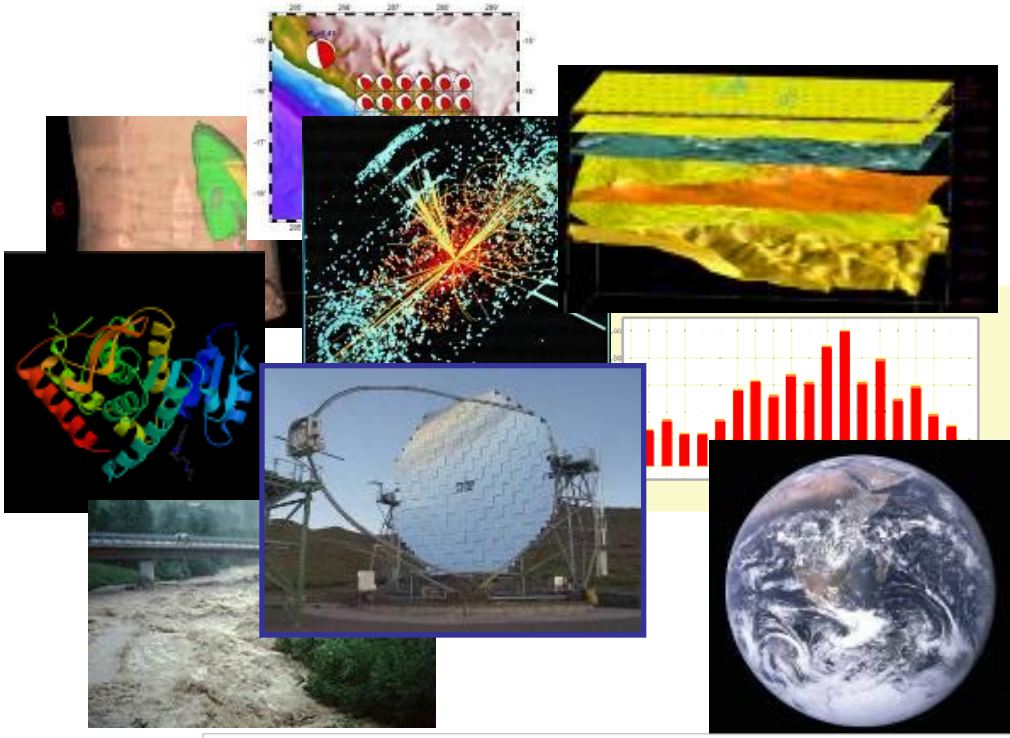


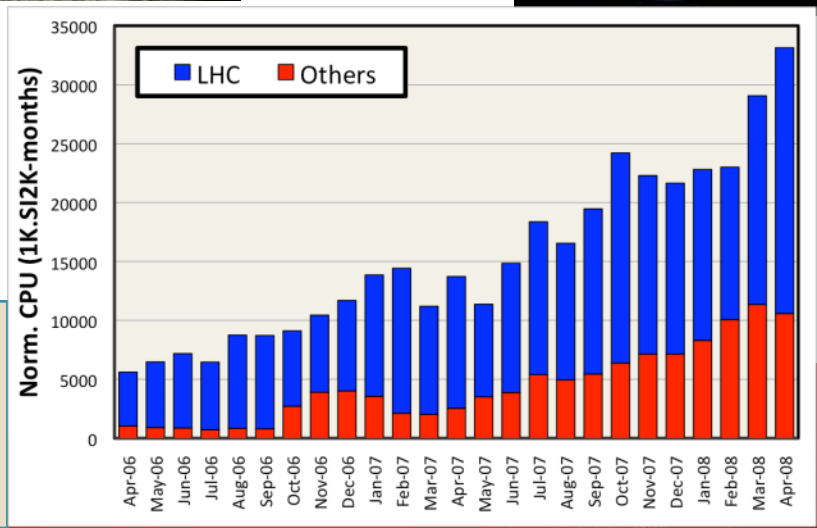
David Groep
Nikhef
Amsterdam
PDP programme

http://dashb-atlas-ddm.cern.ch/ddm2

Shared e-Infrastructure

- >270 communities from many different domains
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High Energy Physics
  - Life Sciences
  - …





David Groep
Nikhef
Amsterdam

Applications have moved from testing to routine and daily usage

~80-95% efficiency

Global data flows

~150GByte, 12hrs
per (human) genome
per sequencer

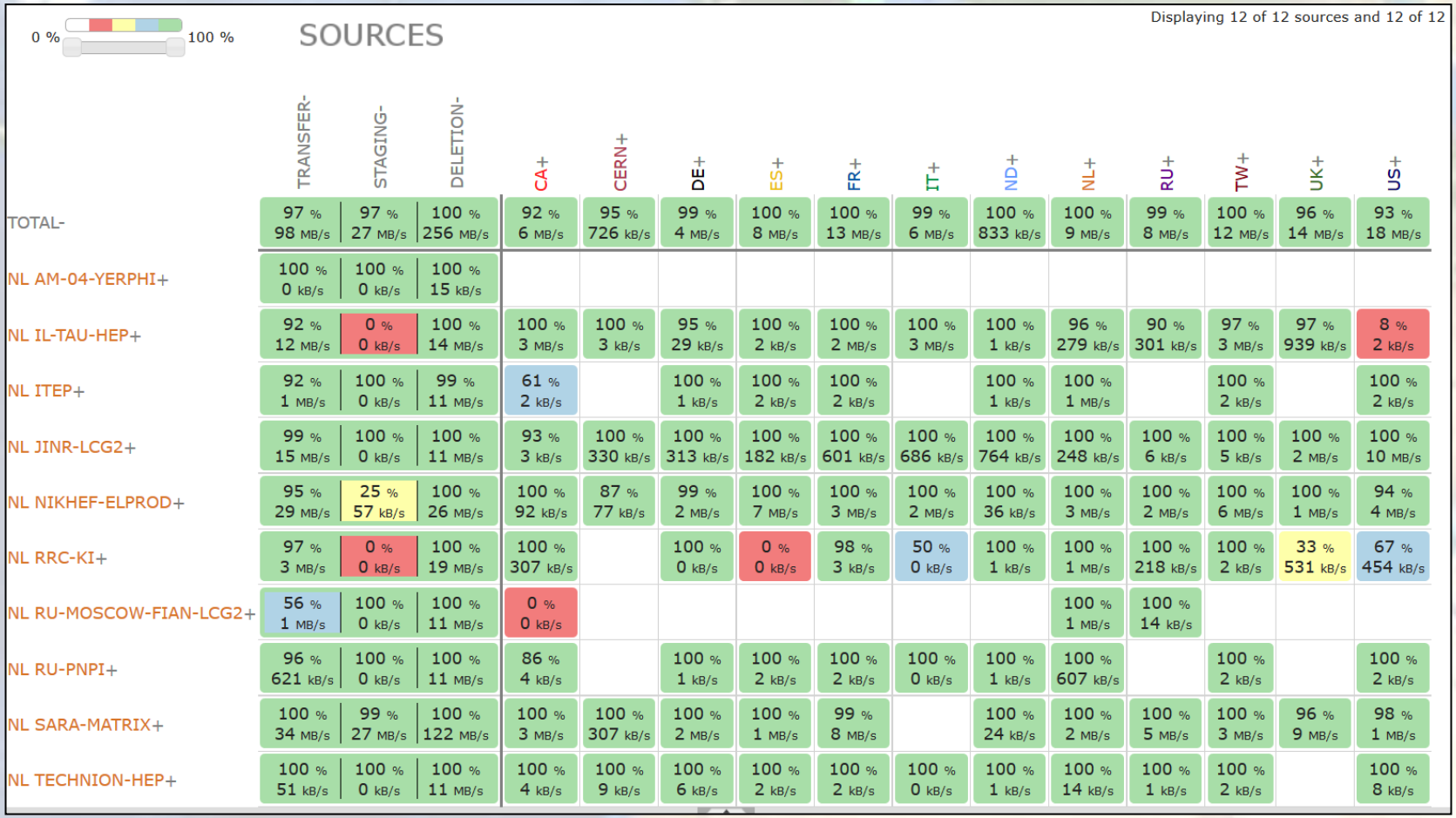
But 1000+ sequencers…

50TByte from Shenzhen to NL is (still) done by rucksack

David Groep
Nikhef
Amsterdam
PDP programme

Genome sequencing at the Beijing Genomics Institute BGI
Photo: Scotted400, CC-BY-3.0

Distributed analysis – ‘Atlas neighbours’



0 % 100 %

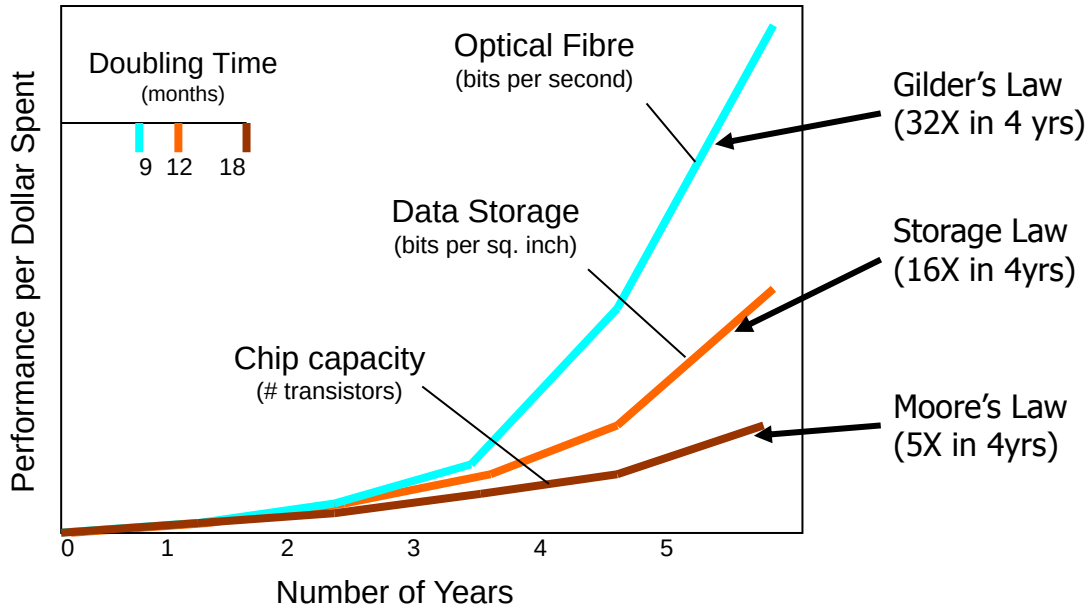
SOURCES

Displaying 12 of 12 sources and 12 of 12

	TRANSFER-	STAGING-	DELETION-	CA+	CERN+	DE+	ES+	FR+	IT+	ND+	NL+	RU+	TW+	UK+	US+
TOTAL-	97 % 98 MB/s	97 % 27 MB/s	100 % 256 MB/s	92 % 6 MB/s	95 % 726 kB/s	99 % 4 MB/s	100 % 8 MB/s	100 % 13 MB/s	99 % 6 MB/s	100 % 833 kB/s	100 % 9 MB/s	99 % 8 MB/s	100 % 12 MB/s	96 % 14 MB/s	93 % 18 MB/s
NL AM-04-YERPHI+	100 % 0 kB/s	100 % 0 kB/s	100 % 15 kB/s												
NL IL-TAU-HEP+	92 % 12 MB/s	0 % 0 kB/s	100 % 14 MB/s	100 % 3 MB/s	100 % 3 kB/s	95 % 29 kB/s	100 % 2 kB/s	100 % 2 MB/s	100 % 3 MB/s	100 % 1 kB/s	96 % 279 kB/s	90 % 301 kB/s	97 % 3 MB/s	97 % 939 kB/s	8 % 2 kB/s
NL ITEP+	92 % 1 MB/s	100 % 0 kB/s	99 % 11 MB/s	61 % 2 kB/s		100 % 1 kB/s	100 % 2 kB/s	100 % 2 MB/s		100 % 1 kB/s	100 % 1 MB/s		100 % 2 kB/s		100 % 2 kB/s
NL JINR-LCG2+	99 % 15 MB/s	100 % 0 kB/s	100 % 11 MB/s	93 % 3 kB/s	100 % 330 kB/s	100 % 313 kB/s	100 % 182 kB/s	100 % 601 kB/s	100 % 686 kB/s	100 % 764 kB/s	100 % 248 kB/s	100 % 6 kB/s	100 % 5 kB/s	100 % 2 MB/s	100 % 10 MB/s
NL NIKHEF-ELPROD+	95 % 29 MB/s	25 % 57 kB/s	100 % 26 MB/s	100 % 92 kB/s	87 % 77 kB/s	99 % 2 MB/s	100 % 7 MB/s	100 % 3 MB/s	100 % 2 MB/s	100 % 36 kB/s	100 % 3 MB/s	100 % 2 MB/s	100 % 6 MB/s	100 % 1 MB/s	94 % 4 MB/s
NL RRC-KI+	97 % 3 MB/s	0 % 0 kB/s	100 % 19 MB/s	100 % 307 kB/s		100 % 0 kB/s	0 % 0 kB/s	98 % 3 kB/s	50 % 0 kB/s	100 % 1 kB/s	100 % 1 MB/s	100 % 218 kB/s	100 % 2 kB/s	33 % 531 kB/s	67 % 454 kB/s
NL RU-MOSCOW-FIAN-LCG2+	56 % 1 MB/s	100 % 0 kB/s	100 % 11 MB/s	0 % 0 kB/s							100 % 1 MB/s	100 % 14 kB/s			
NL RU-PNPI+	96 % 621 kB/s	100 % 0 kB/s	100 % 11 MB/s	86 % 4 kB/s		100 % 1 kB/s	100 % 2 kB/s	100 % 2 kB/s	100 % 0 kB/s	100 % 1 kB/s	100 % 607 kB/s		100 % 2 kB/s		100 % 2 kB/s
NL SARA-MATRIX+	100 % 34 MB/s	99 % 27 MB/s	100 % 122 MB/s	100 % 3 MB/s	100 % 307 kB/s	100 % 2 MB/s	100 % 1 MB/s	99 % 8 MB/s		100 % 24 kB/s	100 % 2 MB/s	100 % 5 MB/s	100 % 3 MB/s	96 % 9 MB/s	98 % 1 MB/s
NL TECHNION-HEP+	100 % 51 kB/s	100 % 0 kB/s	100 % 11 MB/s	100 % 4 kB/s	100 % 9 kB/s	100 % 6 kB/s	100 % 2 kB/s	100 % 2 kB/s	100 % 0 kB/s	100 % 1 kB/s	100 % 14 kB/s	100 % 1 kB/s	100 % 2 kB/s		100 % 8 kB/s

http://dashb-atlas-ddm.cern.ch/ddm2

There's always a network close to you



NL Light

SURF NET

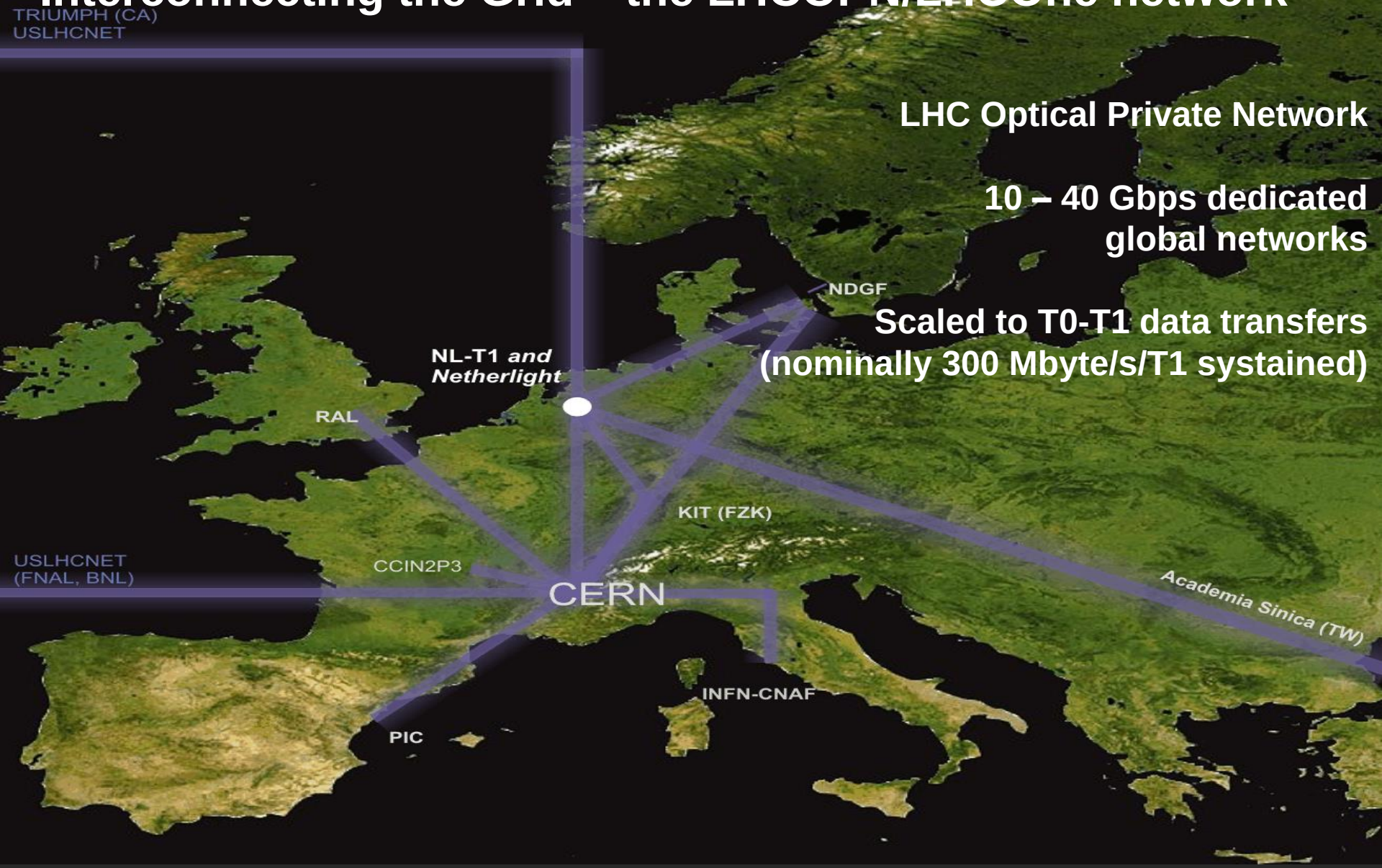
SURFnet pioneered 'lambda' and hybrid networks in the world

- and likely contributed to the creation of a market for 'dark fibre' in the Netherlands

There's always fibre within 2 miles from you – where ever you are!

(it's just that last mile to your home that's missing – and the business model of your telecom provider…)

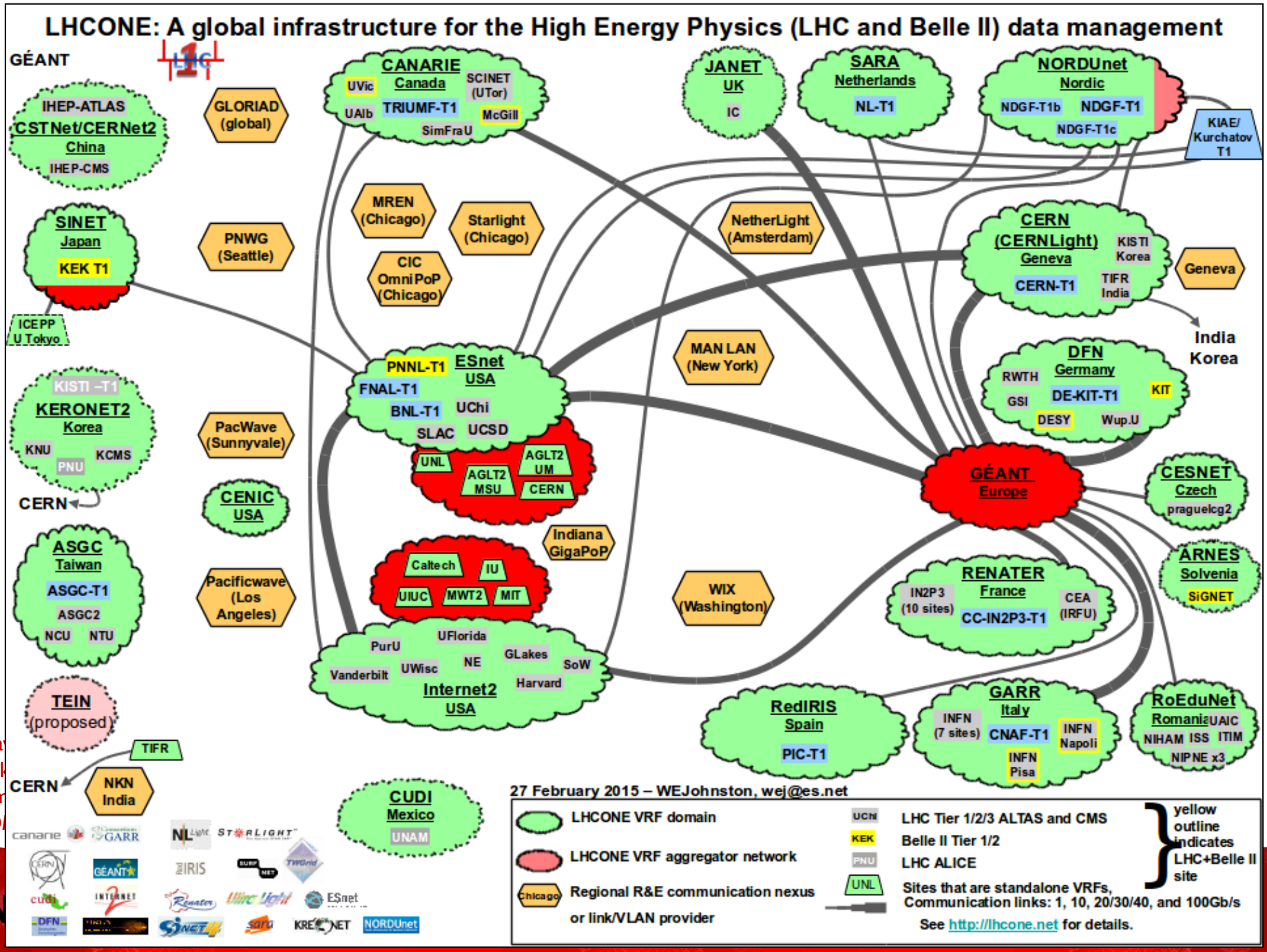
Interconnecting the Grid – the LHCOPN/LHCOne network



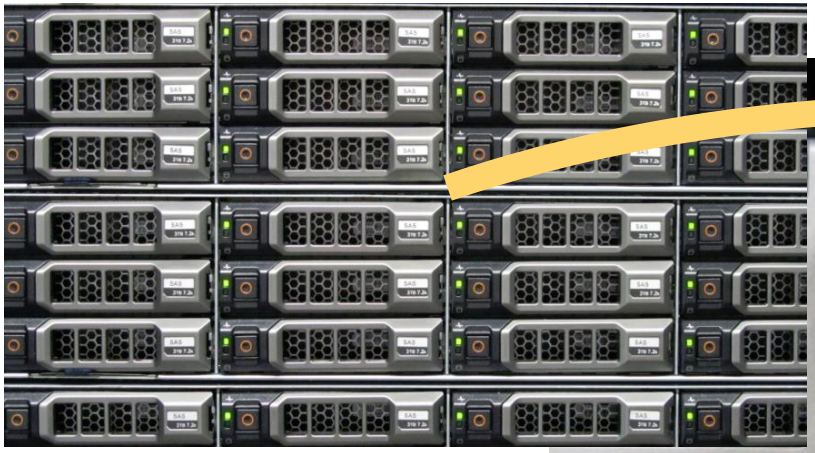
LHC Optical Private Network

10 – 40 Gbps dedicated global networks

Scaled to T0-T1 data transfers (nominally 300 Mbyte/s/T1 systained)


LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management
GÉANT
IHEP-ATLAS
CSTNet/CERNet2
China
IHEP-CMS
GLORIAD (global)
CANARIE
Canada
UVic
SCINET (UTor)
UAlb
TRIUMF-T1
McGill
SimFraU
JANET
UK
IC
SARA
Netherlands
NL-T1
NORDUnet
Nordic
NDGF-T1b
NDGF-T1
NDGF-T1c
KIAE/ Kurchatov T1
SINET
Japan
KEK T1
PNWG (Seattle)
MREN (Chicago)
Starlight (Chicago)
CIC OmniPoP (Chicago)
NetherLight (Amsterdam)
CERN (CERNLight)
Geneva
CERN-T1
KISTI Korea
TIFR India
Geneva
India
Korea
ICEPP
U Tokyo
MAN LAN (New York)
DFN
Germany
RWTH
DE-KIT-T1
KIT
GSI
DESY
Wup.U
KISTI –T1
KERONET2
Korea
KNU
PNU
KCMS
CERN
PacWave (Sunnyvale)
PNNL-T1
ESnet
USA
FNAL-T1
BNL-T1
UChi
SLAC
UCSD
UNL
AGLT2 MSU
AGLT2 UM
CERN
GÉANT
Europe
CESNET
Czech
praguelcg2
CENIC
USA
Indiana GigaPoP
ASGC
Taiwan
ASGC-T1
ASGC2
NCU
NTU
Pacificwave (Los Angeles)
Caltech
IU
UIUC
MWT2
MIT
WIX (Washington)
RENATER
France
IN2P3 (10 sites)
CC-IN2P3-T1
CEA (IRFU)
ARNES
Solvenia
SiGNET
PurU
UFlorida
NE
GLakes
SoW
Vanderbilt
UWisc
Harvard
Internet2
USA
TEIN
(proposed)
TIFR
CERN
NKN
India
RedIRIS
Spain
PIC-T1
GARR
Italy
INFN (7 sites)
CNAF-T1
INFN Napoli
INFN Pisa
RoEduNet
Romania
UAIC
NIHAM
ISS
ITIM
NIPNE x3
CUDI
Mexico
UNAM
27 February 2015 – WEJohnston, wej@es.net
LHCONE VRF domain
LHCONE VRF aggregator network
Chicago
Regional R&E communication nexus or link/VLAN provider
UChi
LHC Tier 1/2/3 ALTAS and CMS
KEK
Belle II Tier 1/2
PNU
LHC ALICE
UNL
Sites that are standalone VRFs, Communication links: 1, 10, 20/30/40, and 100Gb/s
yellow outline indicates LHC+Belle II site
See http://lhcone.net for details.

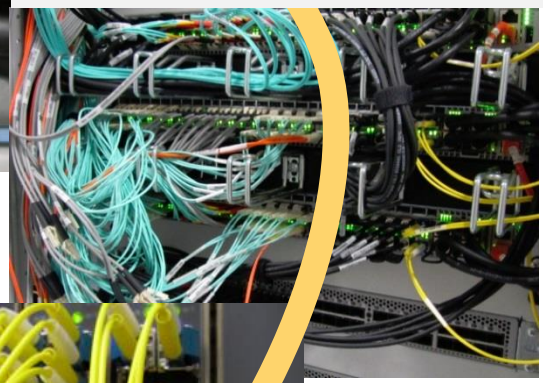
The Flow of Data at Nikhef



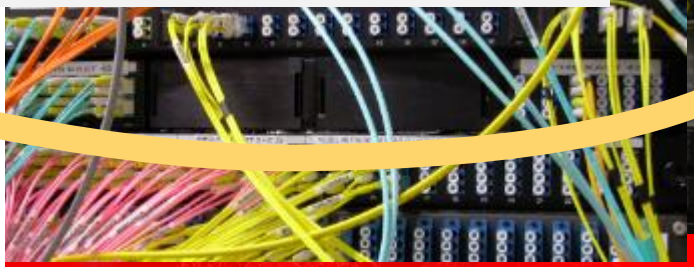
44 disk servers
~3 PiB (~3000 TByte)
2 control & DB nodes

10 – 40 Gbps per server

240 Gbps interconnect

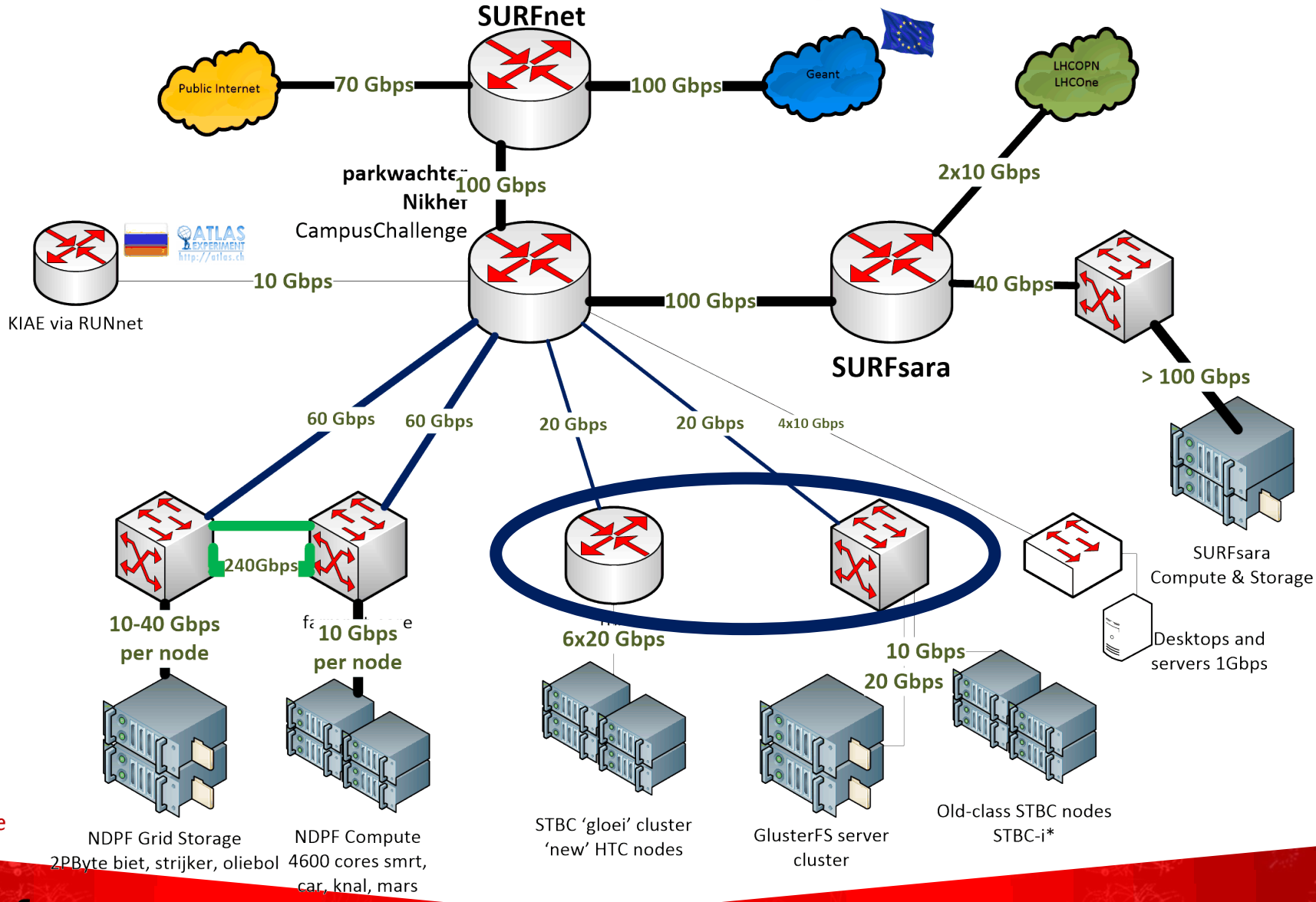


Peerings: SURFnet, SURFsara, Kurchatov, AMOLF, CWI, LHCOPN, LHCOne via SARA



>200 Gbps uplinks

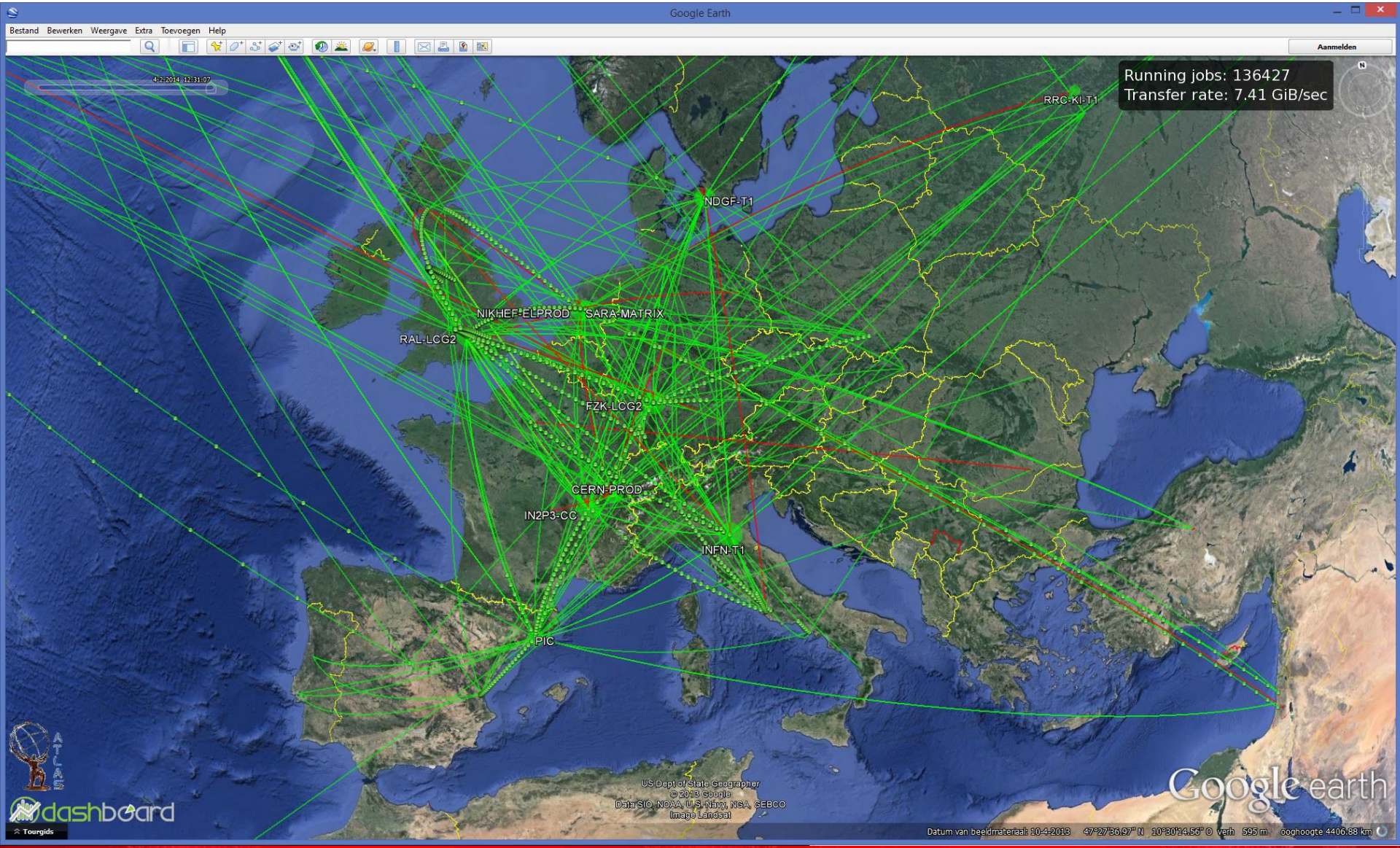
Nikhef Data Processing network



David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

Data Flows in the LHC Computing Grid



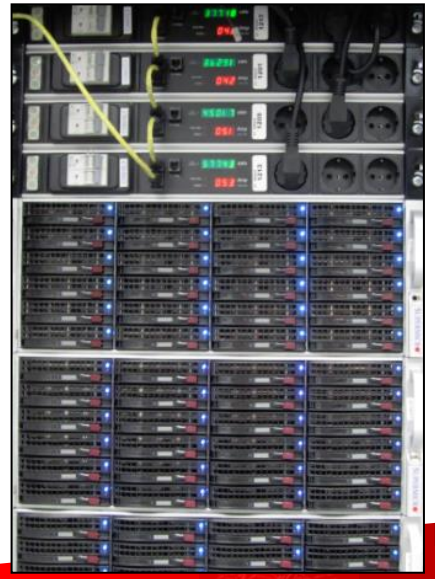
But that's only a small subset

LCG "FTS" visualisation sees but part of the data

- only shows the centrally managed data transfers
- sees only traffic from Atlas, CMS, and LHCb
- cannot show the quality, nor bandwidth used

But each of our nodes sees all its transfers

- server logging is in itself data
- we collect it all



David Groep
Nikhef
Amsterdam
PDP programme

One day worth of logs … ~12GB/day

```
tbn18.nikhef.nl:/var/log/
 631M    dpm/log.1
 1.1G    dpns/log.1
 639M    srmv2.2/log.1
```

plus 44 disk server nodes @250 Mbyte/day

```
09/13 00:01:23.588 26759,79 dpm_srv_proc_put: calling dpm_selectfs
09/13 00:01:23.588 26759,79 dpm_selectfs: selected pool: BIOMED
09/13 00:01:23.588 26759,79 dpm_selectfs: selected file system: oliebol-02.nikhef.nl:/export/data/biomed
09/13 00:01:23.588 26759,79 dpm_selectfs: oliebol-02.nikhef.nl:/export/data/biomed reqsize=0, elemp->free=399976081749, poolp->free=399976081749
09/13 00:01:23.645 26759,79 dpm_srv_proc_put: calling Cns_creatx
09/13 00:01:23.712 26759,19 dpm_srv_getspacetoken: DP092 - getspacetoken request by /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=53149
09/13 00:01:23.712 26759,19 dpm_srv_getspacetoken: DP098 - getspacetoken ATLASDATADISK
09/13 00:01:23.713 26759,19 dpm_srv_getspacetoken: returns 0, status=DPM_SUCCESS
09/13 00:01:23.753 26759,78 dpm_srv_proc_get: TURL info: gsiftp oliebol-09.nikhef.nl oliebol-09.nikhef.nl:/export/data/atlasprd/atlas/2015-09-12/
09/13 00:01:23.755 26759,78 dpm_srv_proc_get: returns 0, status=DPM_SUCCESS
09/13 00:01:23.761 26759,79 dpm_srv_proc_put: TURL info: gsiftp oliebol-02.nikhef.nl oliebol-02.nikhef.nl:/export/data/biomed/biomed/2015-09-13/j
09/13 00:01:23.763 26759,79 dpm_srv_proc_put: returns 0, status=DPM_SUCCESS
09/13 00:01:23.881 26759,18 dpm_updfreespace: oliebol-02.nikhef.nl:/export/data/biomed incr=0, elemp->free=399976081749, poolp->free=399976081749
09/13 00:01:23.881 26759,18 dpm_srv_putdone: returns 0, status=DPM_SUCCESS
```

David Groep
Nikhef
Amsterdam
PDP programme

And then our storage manager is still ‘decent’ …

CASTOR Logs

- ~2GB/day from the node I showed (highest volume)
- ~30GB/day collected overall
- ~200 source nodes
- ~70,000,000 log events/day

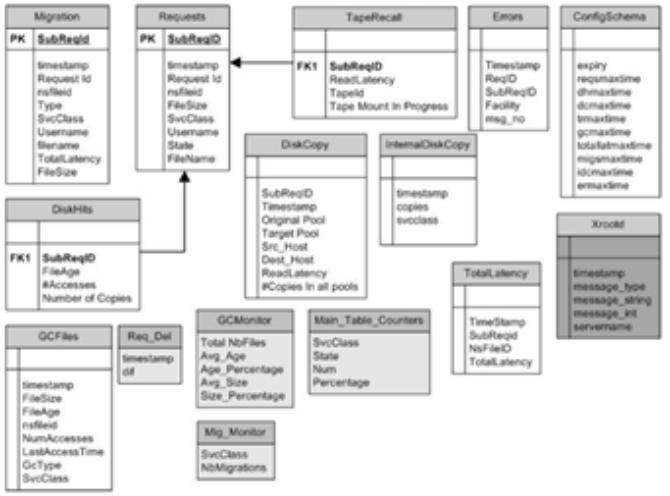
David Groep
Nikhef
Amsterdam
PDP programme

Science & Technology Facilities Council

Nikhef

The First Solution - DLF

- DLF = 'Distributed Logging Facility'
- CERN-developed monitoring system for CASTOR
- Store all the log information in a big Oracle DB



Source: *CASTOR end-to-end monitoring*, by T Rekatsinas et al, URL: http://iopscience.iop.org/1742-6596/219/4/042052/pdf/1742-6596_219_4_042052.pdf

David Groep
Nikhef
Amsterdam
PDP programme

Science & Technology Facilities Council

Running DLF

- Scalability was a killer.
  - By 2013, simple queries were taking >1 hour.
  - Fundamental architecture couldn't cope.

David Groep
Nikhef
Amsterdam
PDP programme

Science & Technology Facilities Council

Nikhef

Big Data Analytics for log analysis

- Analysis of log data is typical 'big data' problem
  - CERN tried Hadoop ('map-reduce')
  - RAL went with … ELK*
- For logs specifically, it's mostly efficient search
  - ElasticSearch (www.elastic.co)
  - LogStash (collect and parse logs, import to ES)
  - Kibana – analysis based on Apache Lucene + graphing
- Integrated into a single 'stack': ELK

David Groep
Nikhef
Amsterdam
PDP programme

*Appleyard et al., http://pos.sissa.it/archive/conferences/239/027/ISGC2015_027.pdf

LogStash

Data arrives in format A…
…process B occurs…
…data out in format C

e.g. convert syslog into json

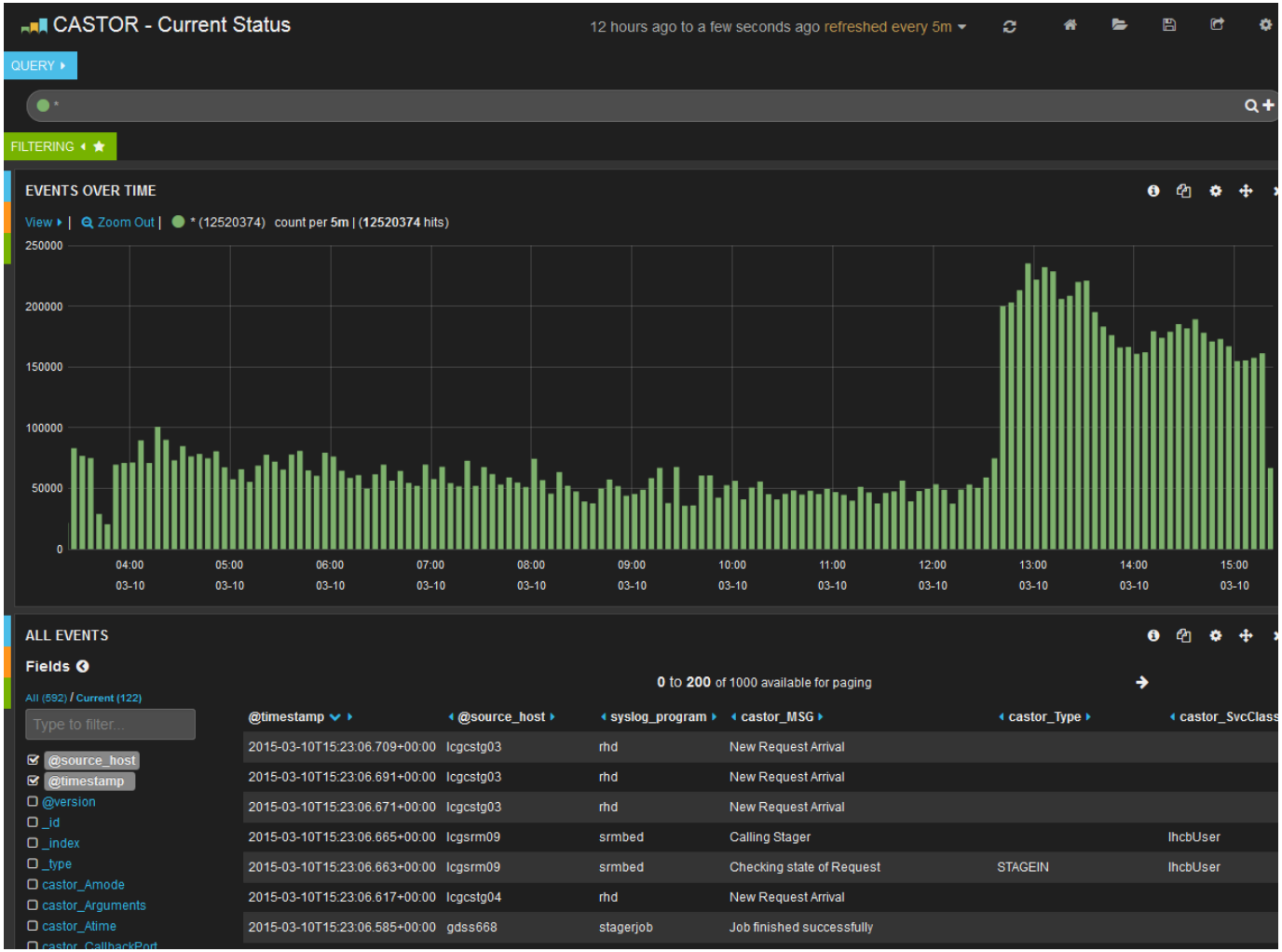
plugin documentation

inputs	codecs	filters	outputs
collectd	cloudtrail	advisor	boundary
drupal_dblog	collectd	alter	circonus
elasticsearch	compress_spooler	anonymize	cloudwatch
eventlog	dots	checksum	csv
exec	edn	cidr	datadog
file	edn_lines	cipher	datadog_metrics
ganglia	fluent	clone	elasticsearch
gelf	graphite	collate	elasticsearch_http
gemfire	json	csv	elasticsearch_river
generator	json_lines	date	email
graphite	json_spooler	dns	exec
heroku	line	drop	file
imap	msgpack	elapsed	ganglia
invalid_input	multiline	elasticsearch	gelf
irc	netflow	environment	gemfire
jmx	noop	extractnumbers	google_bigquery
log4j	oldlogstashjson	fingerprint	google_cloud_storage
lumberjack	plain	gelfify	graphite
pipe	rubydebug	geoip	graphtastic
puppet_facter	spool	grep	hipchat
rabbitmq		grok	http
rackspace		grokdiscovery	irc
redis		i18n	jira
relp		json	juggernaut
s3		json_encode	librato
snmptrap		kv	loggly
sqlite		metaevent	lumberjack

David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

Analyse ‘by hand’: Kibana & Lucene



David Groep
Nikhef
Amsterdam
PDP programme

Graphic from Rob Appleyard, RTFC RAL – at ISGC 2015

Challenges ahead!

At the start

- 44+ difference data sources, 240 Gbps of traffic
- 150+ different storage partners, 55 countries/regions
- public internet plus the LHCOPN/LHCOne
- 5000+ users, working 24x7

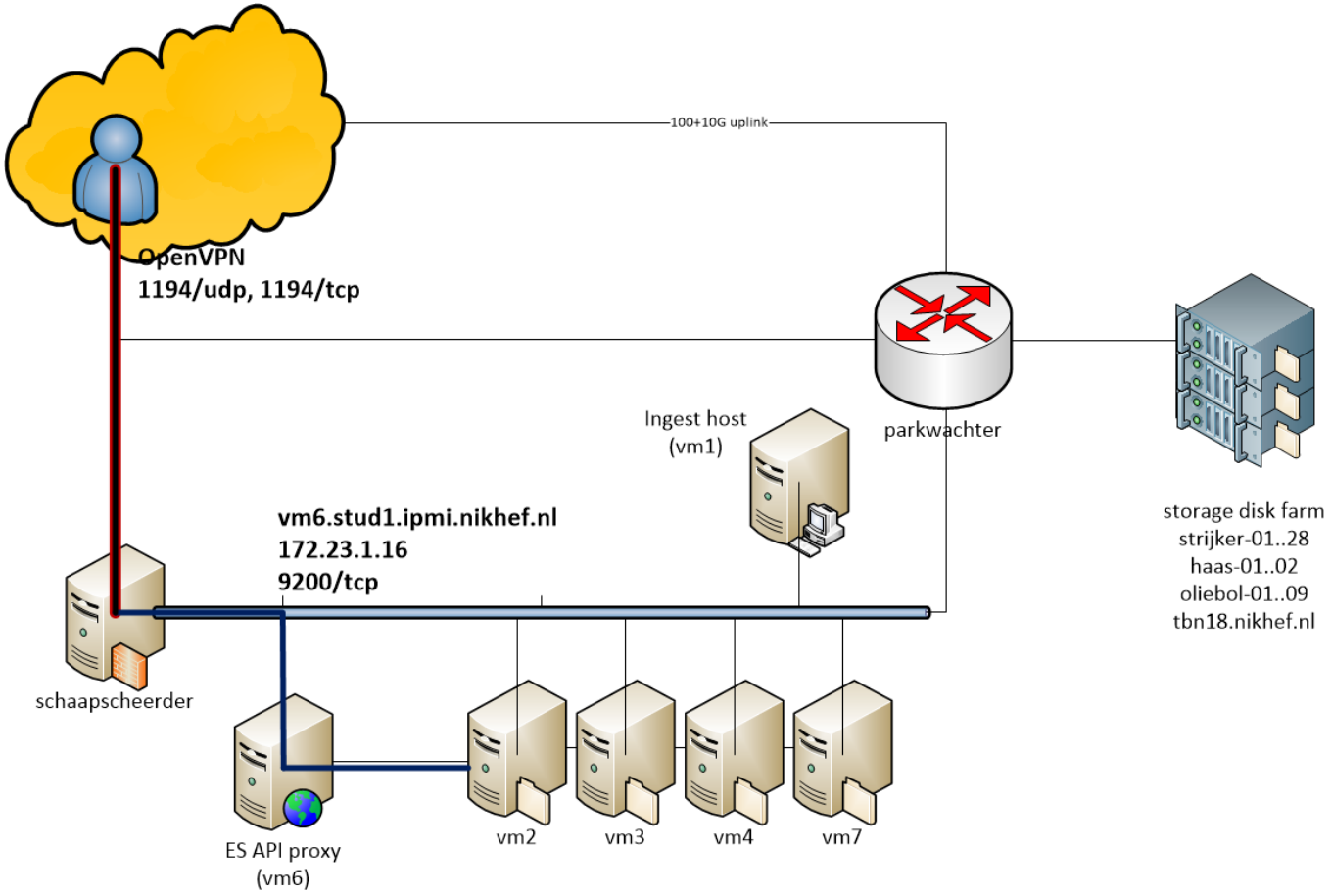
… and 'grep' for a tool … ☹

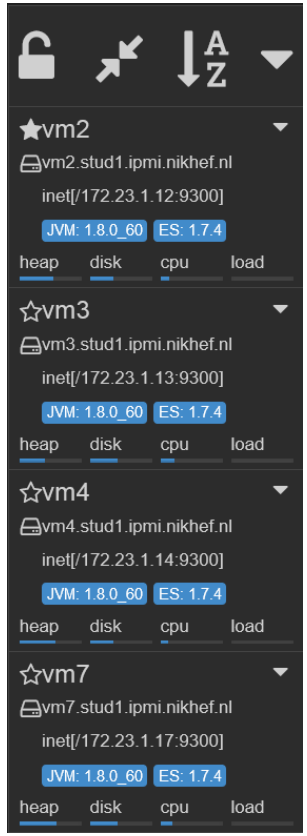
And phase I added

- setup of a big data analytics cluster (ELK)
- merge diverse data sources into a single system
- define queries and find some global anomalies ☺

David Groep
Nikhef
Amsterdam
PDP programme

ELK Cluster and Interface





David Groep
Nikhef
Amsterdam
PDP programme

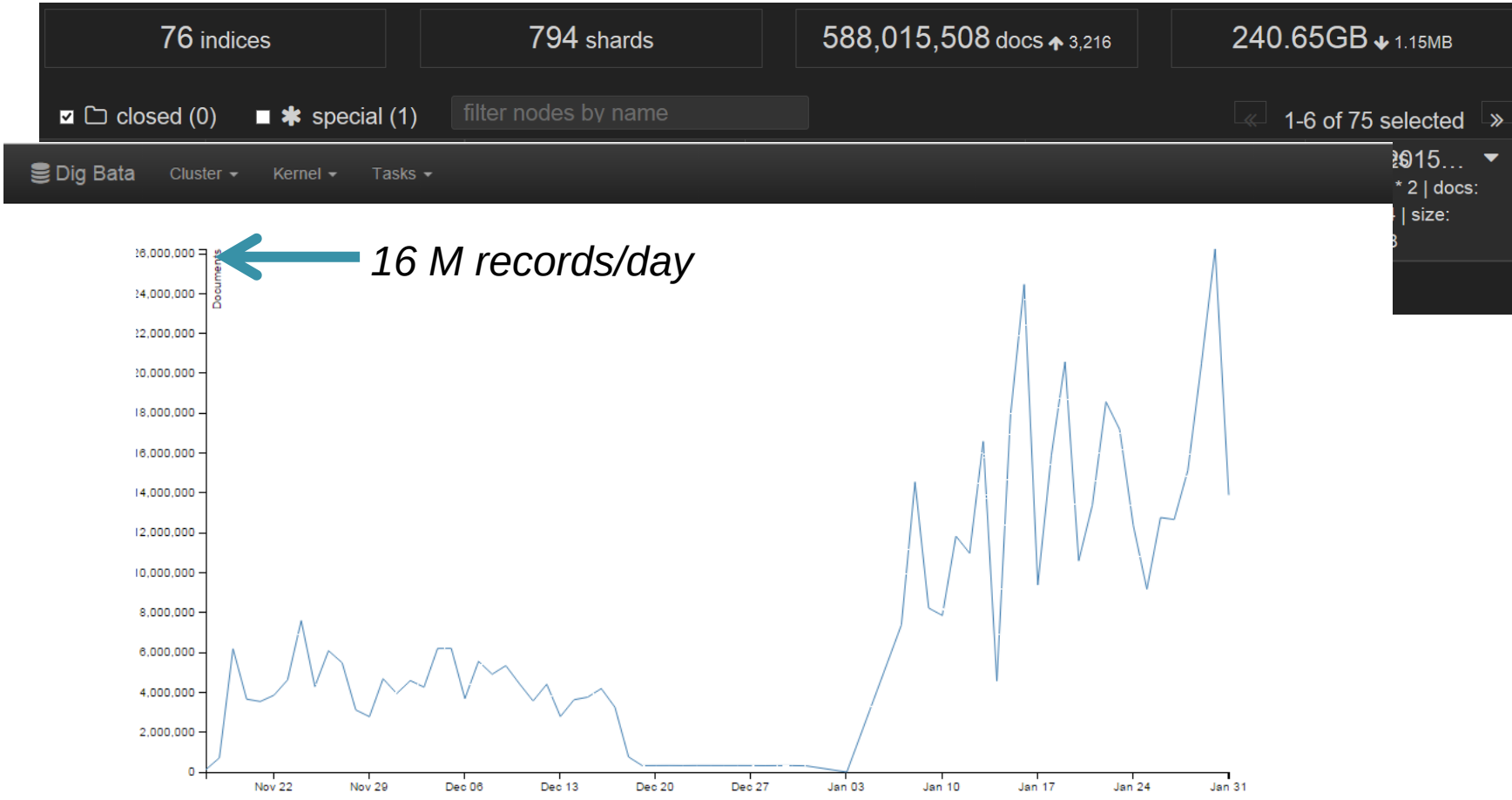
Filled with data

filter nodes by name | master | data | client

name	load average	cpu %	heap usage %	disk usage %	uptime
logstash-vm5.stud1.ipmi.nikhef.nl-7420-... vm5.stud1.ipmi.nikhef.nl inet[/172.23.1.15:9300] JVM: 1.8.0_60 ES: 1.7.0	N/A		25.0 used: 126.69MB max: 491.69MB	*no disk info for client nodes*	3d.
vm2 vm2.stud1.ipmi.nikhef.nl inet[/172.23.1.12:9300] JVM: 1.8.0_60 ES: 1.7.4		22.0	59.0 used: 2.36GB max: 3.98GB	38.0 free: 121.55GB total: 196.74GB	9d.
vm3 vm3.stud1.ipmi.nikhef.nl inet[/172.23.1.13:9300] JVM: 1.8.0_60 ES: 1.7.4		18.0	58.0 used: 2.33GB max: 3.98GB	45.0 free: 107.77GB total: 196.74GB	7d.
vm4 vm4.stud1.ipmi.nikhef.nl inet[/172.23.1.14:9300] JVM: 1.8.0_60 ES: 1.7.4		15.0	73.0 used: 2.94GB max: 3.98GB	37.0 free: 123.51GB total: 196.74GB	7d.
vm7 vm7.stud1.ipmi.nikhef.nl inet[/172.23.1.17:9300] JVM: 1.8.0_60 ES: 1.7.4		20.0	59.0 used: 2.36GB max: 3.98GB	44.0 free: 109.89GB total: 196.74GB	9d.

David Groep
Nikhef
Amsterdam
PDP programme

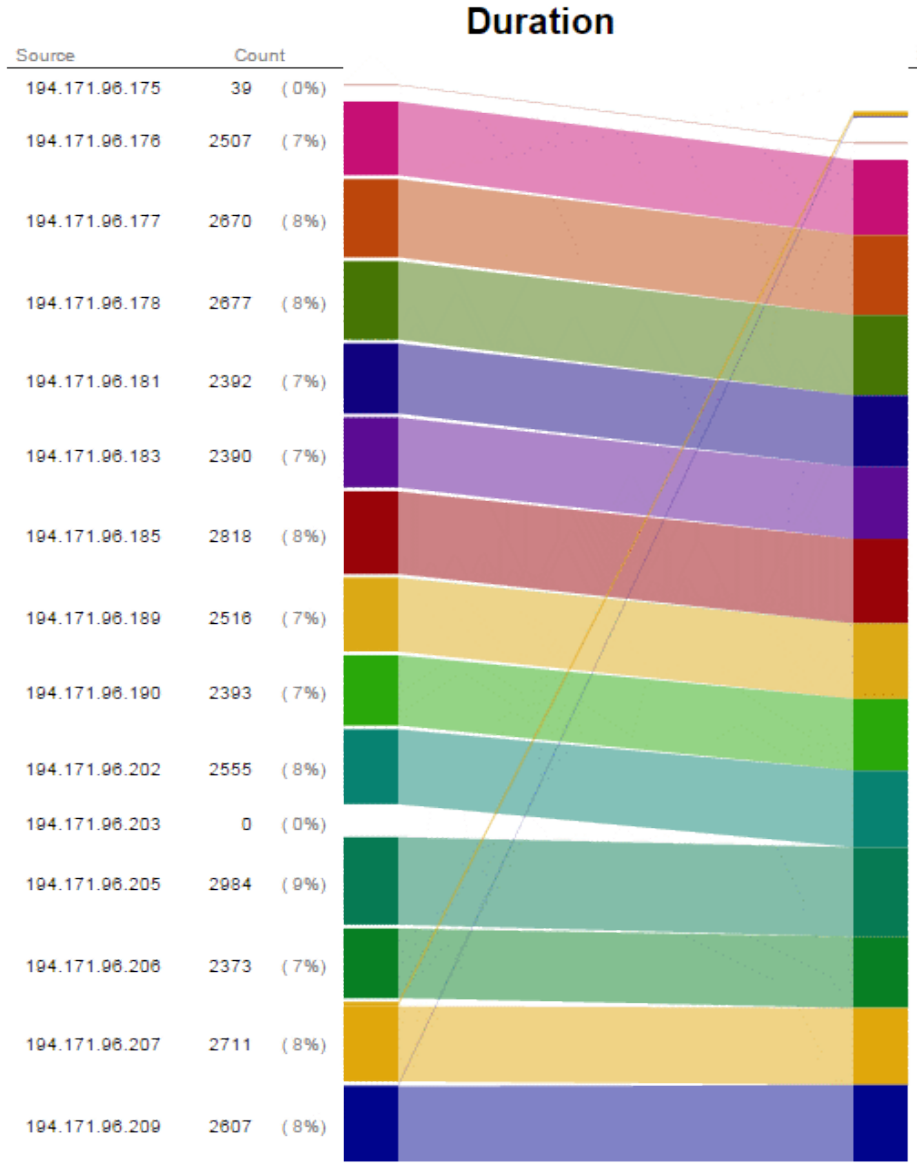
Filling it up: from 0 to ~ 500 million

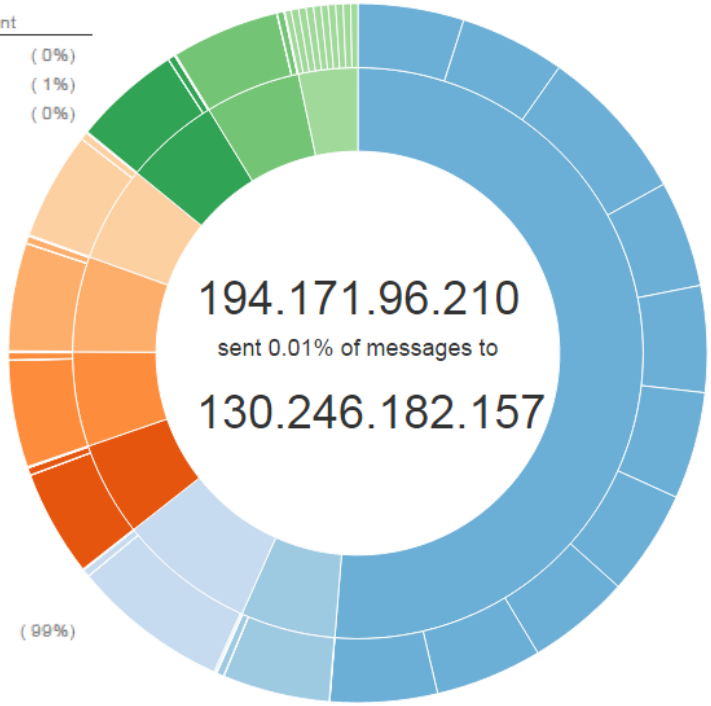


David Groep
Nikhef
Amsterdam
PDP programme

Graphics courtesy Jouke Roorda, Olivier Verbeek, Rens Visser

Nikhef





Data flows: a user at UC Irvine reading a data set from Nikhef, with some background from CERN and KIT Karlsruhe

Graphics courtesy Jouke Roorda, Olivier Verbeek, Rens Visser

Nikhef

Building upon phase I …

Phase II

- How can global data flows be presented?
- Can one conceive visualisations for the general public? for users? or for both?
- Identifying troublesome inter-peer links (and local disk servers) via analytics techniques
- Identifying 'hot' (interesting) data by analysing usage
- Pro-active security incident detection:
  are there users 'behaving strangely'?

David Groep
Nikhef
Amsterdam
PDP programme

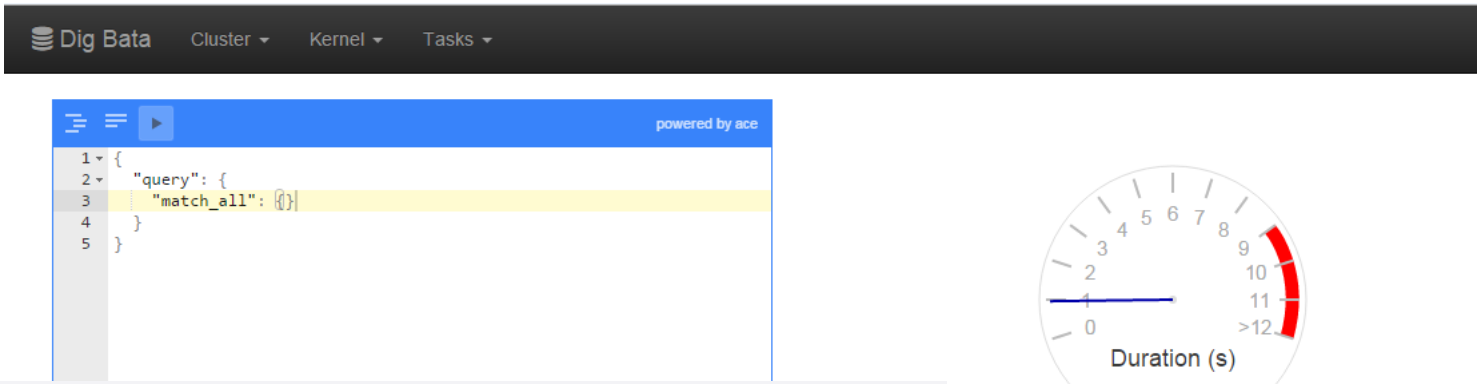
Phase II: plenty to do!

…

- What's the way to explain WLCG data flows to the world?
- which systems (or group of systems) uses the most bandwidth (in and out separately)
- which systems (or group of systems) generates the most connections?
- what does the spectrum of transfers look like? mostly big, mostly small, ... ?
- what does the time distribution of transfers look like (bandwidth and also number)?

David Groep
Nikhef
Amsterdam
PDP programme

Queries and responses




```
{
    "Query_begin_structuur": {
        "type_query": {
            "field": "naam field aan te spreken"
        }
    }
}
```

David Groep
Nikhef
Amsterdam
PDP programme

Graphics courtesy Jouke Roorda, Olivier Verbeek, Rens Visser

Nikhef

Access services

https://wiki.nikhef.nl/grid/HvABigDataVisualisation

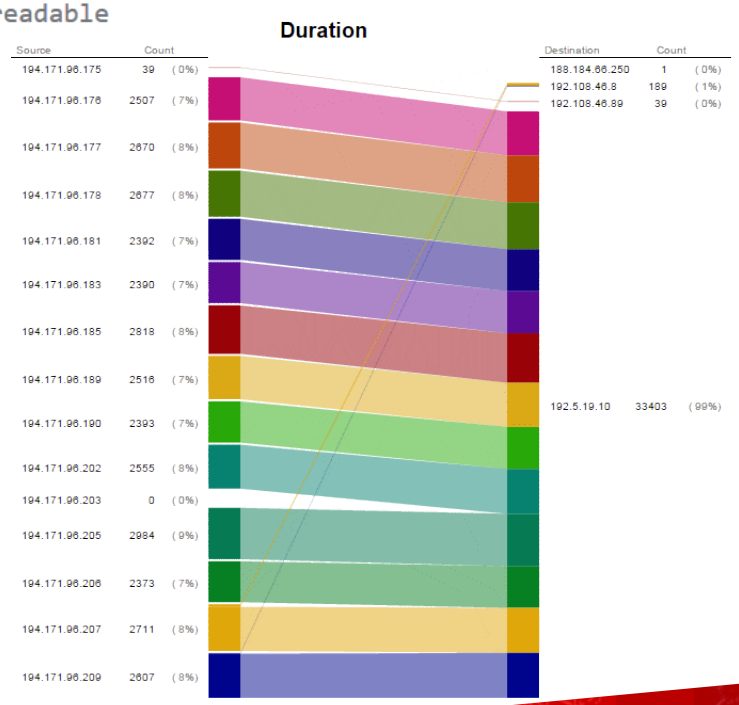
- OpenVPN configuration
- Access to full ES search API

- Please limit yourself to query methods
- Do not redistribute or copy records with PII
- Kindly observe the Nikhef AUP as well – and report any incidents to security@nikhef.nl

David Groep
Nikhef
Amsterdam
PDP programme

```
public List getDocsHistogram() throws IOException {
    return client.execute(
            new Search.Builder(new SearchSourceBuilder()
                    // This query searches for the right kind of log messages
                    .query(
                            QueryBuilders.boolQuery()
                                    .must(QueryBuilders.matchQuery("program", "dpm-gsiftp"))
                                    .must(QueryBuilders.matchQuery("msg_type", "transfer"))
                    )
                    // Newest results should be retrieved
                    .sort(SortBuilders.fieldSort("@timestamp").order(SortOrder.DESC))
                    // Get 25 entries to keep the graph readable
                    .size(25)
                    .toString()
            ).addIndex(logsIndex).build())
            .getHits(Map.class)
            // Java 8 streams because they're awesome :)
            .stream()
```

Your call!
Questions?



David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

Code sample: Jouke Roorda, PDP-sample-app